DESIGN OF NEAR-FIELD CODED APERTURE CAMERAS FOR HIGH-RESOLUTION MEDICAL AND INDUSTRIAL GAMMA-RAY IMAGING

by

Roberto Accorsi

Dottore in Ingegneria Nucleare, Politecnico di Milano, 1996
S.M., Massachusetts Institute of Technology, 1998

SUBMITTED TO THE DEPARTMENT OF NUCLEAR ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN NUCLEAR ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2001

Signature of Author _____
Department of Nuclear Engineering
May 1, 2001

Certified by _____
Richard C. Lanza
Senior Research Scientist, Department of Nuclear Engineering
Thesis Advisor

Certified by _____
Berthold K. P. Horn
Professor of Electrical Engineering and Computer Science
Thesis Reader

Accepted by _____
Professor Sow-Hsin Chen
Chairman, Department Committee on Graduate Students

Department of Nuclear Engineering

DESIGN OF NEAR-FIELD CODED APERTURE CAMERAS FOR HIGH-RESOLUTION MEDICAL AND INDUSTRIAL GAMMA-RAY IMAGING

by

Roberto Accorsi

Submitted to the Department of Nuclear Engineering
on May 1, 2001 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Nuclear Engineering

ABSTRACT

Coded Aperture Imaging is a technique originally developed for X-ray astronomy, where typical imaging problems are characterized by far-field geometry and an object made of point sources distributed over a mainly dark background. These conditions provide, respectively, the basis of artifact-free and high Signal-to-Noise Ratio (SNR) imaging.

When the coded apertures successful in far-field problems are used in near-field geometry, images are affected by extensive artifacts. The classic remedy is to move away from the object until a far-field geometry is restored, but this is at the expense of counting efficiency and, thus, of the SNR of the images. It is shown in this thesis that the application to near-field of a technique originally developed to mitigate the effects of non-uniform background in far-field applications results in a considerable reduction of near-field artifacts. This result opens the way to the exploitation in near-field problems of the favorable SNR characteristics of coded apertures: images comparable to those provided by state-of-the-art imagers can be obtained in a shorter time or while administering a lower dose to patients.

Further developments follow when the SNR increase is traded for better resolution at constant time and dose. The main focus of this work is on a coded aperture camera specifically designed for high-resolution single-photon planar imaging with a pre-existing gamma (Anger) camera. Original theoretical findings and the results of computer simulations led to an optimal coded aperture that was tested experimentally in phantom as well as *in-vivo* studies. Results include, but are not limited to, 1.66-mm-resolution images of ^{99m}Tc-labeled blood and bone agents in a mouse. The theoretical bases for extension to sub-millimeter resolution and higher-energy isotopes are also laid and a candidate aperture capable of 0.96-mm resolution proposed. Potential applications are in small-animal imaging, pediatric nuclear medicine and breast imaging, where increased resolution can result in earlier diagnosis of disease.

The last Chapter of the thesis extends the ideas developed to the design of a coded aperture suitable for CAFNA (Coded Aperture Fast Neutron Analysis), a contraband detection technique that has been under development at MIT for a number of years.

Thesis Supervisor: Dr. Richard C. Lanza
Title: Senior Research Scientist, Department of Nuclear Engineering

ACKNOWLEDGEMENTS

When I started looking for a topic for my bachelor's thesis I was hoping to find a project on medical imaging. Unfortunately, this was not possible: it is one of the very few regrets I have about my education at the Politecnico di Milano. When I first came to MIT I hoped that some day I would be able to fulfill this aspiration. My thesis advisor, Dr. Richard Lanza, is he who made it possible. It would be inadequate to thank him only for his experience, guidance and counsel. He patiently allowed me to deviate from the main focus of his research and, on the contrary, selflessly (and, I should add, recklessly!) encouraged me on my own way. He has been all but a remote academic advisor. Over the last three years, almost on a daily basis, he has spent part of his time with me, not only in his office, but especially in the labs. No matter how busy he may have been, his door has never been closed. He always wanted me to participate to conferences and meet people: without them this work would have not been the same.

Professor Horn has carefully read this thesis. His review was so thorough that he wrote and ran his own computer codes to verify my findings.

I would like to thank Dr. Albert Brandenstein and Jim Petrousky of the Office of National Drug Control Policy for their continuous support over the three years of this project.

To Dr. Francesca Gasparini of the Politecnico di Milano must go a good share of the credit for the core ideas of this work. We thought together about patterns, symmetries, signals and variances. This office has never been the same since she left, in productivity, cheerfulness, and esthetics.

I wish to thank some of the people involved in my research. They are Bob Zimmerman, of Brigham and Women's Hospital and the Harvard Medical School; Dawid Schellingerhout and Umar Mahmood of the Center for Molecular Imaging Research of the Massachusetts General Hospital; and Joel Lazewatsky of DuPont Pharmaceuticals. A special thank you goes to Fred Cote and Rocky Albano at the student machine shop of MIT's Edgerton center.

Many other people made this work possible in many different, but equally indispensable, ways.

My landlady, Jane Cawley, is a very special person. Some of my best memories ever are related to her and her home.

I have always felt deeply connected to my friends: Daniele, Count and Marquis de Giuffridà (he also holds many more important titles too long to be mentioned here), Jacopone, i due biechi, Ema and "il Fabio," and, more recently, Andrea and Andrea. Francesca (yes, the same as above!) can not be acknowledged here only as Dr. Gasparini, despite her multiple attempts at killing me. The most welcome was her baking. Eric Empey was a great labmate in the most difficult time of this project.

It is among these friends that I would like to thank Professor Apostolakis and Professor Yip for their support and counsel.

I am indebted to several great teachers and professors I have met in my long career as a student. Even if it may not show, be assured that they outnumbered the bad ones. The undergraduate curriculum at the Politecnico di Milano is second to none. From a less professional perspective, I owe the most to the days at the Liceo Scientifico of San Donato, Milan. One of my biggest fortunes in life is to have met Professor Orlando Mazzetti, who taught me much more than Italian Literature and Latin. My first science teacher ever was my maternal grandfather, to whom this thesis is dedicated.

TABLE OF CONTENTS

ABSTRACT 3

ACKNOWLEDGEMENTS 5

TABLE OF CONTENTS 6

LIST OF SYMBOLS 10

LIST OF ACRONYMS 12

INTRODUCTION 13

PART I: INTRODUCTION AND BACKGROUND 17

CHAPTER 1 OVERVIEW AND HISTORICAL BACKGROUND 19

1.1 THE CLASSIC METHODS OF 2D IMAGING 19

1.2 WHY A CODED APERTURE? 21

1.3 CODED APERTURE IMAGING IN A NUTSHELL 22

1.4 CODED APERTURE HISTORY 25

1.5 APPLICATIONS OF CODED APERTURES 26

1.6 THESIS OUTLINE 27

CHAPTER 2 FUNDAMENTALS OF CODED APERTURE IMAGING 29

2.1 THE PINHOLE CAMERA 29

2.2 ENCODING THE SIGNAL: OBJECT PROJECTION 31

2.3 DECODING: COMPUTER POST-PROCESSING 35
2.4 CODED APERTURE FAMILIES 38
2.5 CODED APERTURE CAMERA GEOMETRIES 58
2.6 FIELD OF VIEW AND RESOLUTION 60
2.7 DECODING TECHNIQUES 66
2.8 DEPTH OF FOCUS AND 3D LAMINOGRAPHY 72

PART II: METHODS, THEORETICAL ADVANCEMENTS AND EXPERIMENTAL RESULTS **75**

CHAPTER 3 SIMULATION TOOLS: COMPUTER CODES AND OPTICAL BENCH **77**

3.1 THE SIEMENS E-CAM 77
3.2 THE SIMULATION CODE 78
3.3 THE OPTICAL SIMULATOR 85
3.4 VALIDATION WITH E-CAM DATA 90

CHAPTER 4 THE SIGNAL-TO-NOISE RATIO IN CODED APERTURE IMAGING **93**

4.1 AN INTUITIVE SUMMARY OF THE PROBLEM 94
4.2 SNR DEFINITION AND CHOICE OF THE DECODING COEFFICIENTS 95
4.3 THE SIGNAL-TO-NOISE RATIO OF DIFFERENT CODED APERTURES 103
4.4 COMPARING THE SNR PERFORMANCE OF DIFFERENT ARRAYS 111
4.5 DEPENDENCE OF THE VARIANCE ON OTHER SOURCES 115
4.6 SIMULATIONS 117
4.7 OBSERVATION ON THE RELATION BETWEEN SNR AND SENSITIVITY 117

CHAPTER 5 ARTIFACT THEORY **119**

5.1 MASK TRANSMISSION 119
5.2 SAMPLING 120
5.3 ROTATIONAL MISALIGNMENT 123
5.4 NEAR-FIELD ARTIFACTS 123
5.5 VERIFYING THE NEAR-FIELD ARTIFACT THEORY 133
5.6 NEAR-FIELD ARTIFACT REDUCTION 135

5.7 EXPERIMENTAL RESULTS 136
5.8 MASK THICKNESS ARTIFACTS 136
5.9 SUMMARY 137

CHAPTER 6 DESIGN AND FABRICATION OF A CODED APERTURE: EXPERIMENTAL RESULTS 141

6.1 MASK DESIGN 141
6.2 EXPERIMENTAL RESULTS 147
6.3 ELECTRONIC FOCUSING 150
6.4 IN VIVO EXPERIMENTS 152
6.5 SUMMARY 156

PART III: ADVANCED TOPICS AND FUTURE DEVELOPMENTS 159

CHAPTER 7 ADVANCED MASK DESIGN 161

7.1 RESOLUTION IN A REAL DETECTOR 161
7.2 CHOOSING A CONFIGURATION 170
7.3 THE DESIGN OF AN OPTIMAL RESOLUTION MASK 171
7.4 AN IMPROVED FIGURE OF MERIT 174
7.5 EXPERIMENTAL RESULTS: ^{111}IN 182
7.6 SUMMARY 184

CHAPTER 8 EXTENSIONS AND FUTURE WORK 187

8.1 EXPERIMENTS WITH THE CURRENT MASK 187
8.2 EXTENSION TO HIGHER ENERGY 188
8.3 ARTIFACT REDUCTION 191
8.4 THREE-DIMENSIONAL IMAGING 191

CHAPTER 9 APPLICATION TO CAFNA 197

9.1 BASIC PRINCIPLES OF CAFNA 197
9.2 THE BASICS OF FAST NEUTRON ANALYSIS 198
9.3 A 1D CARBON IMAGE 211

9.4 DESIGN OF A CODE APERTURE FOR CAFNA 214
9.5 SUMMARY 215

CHAPTER 10 CONCLUSIONS **217**

APPENDICES **221**

BIBLIOGRAPHY **251**

LIST OF SYMBOLS

CAPITAL BOLDFACE indicates a discrete 2d function or matrix.

Lowercase boldface indicates a 1d array.

CAPITAL ITALIC indicates a continuous 2d function or a scalar variable.

Lowercase italic indicates a continuous 1d function or a scalar variable.

$\equiv$: definition

$\times$: correlation: $H(\vec{y}) = F(\vec{x}) \times G(\vec{x}) = \iint F(\vec{x})G(\vec{y}+\vec{x})d^2\vec{x}$

$\otimes$: correlation (periodic): $H(\vec{y}) = F(\vec{x}) \otimes G(\vec{x}) = \iint F(\vec{x})G(\vec{y} \oplus \vec{x})d^2\vec{x}$

$*$: convolution: $H(\vec{y}) = F(\vec{x}) * G(\vec{x}) = \iint F(\vec{x})G(\vec{y}-\vec{x})d^2\vec{x}$

$\oplus$: sum modulo p

$\mathcal{D}$: cyclic difference set

$\mathfrak{F}$: Fourier transform operator

$\mathfrak{F}^{-1}$: Inverse Fourier transform operator

$\mathfrak{R}$: reflection operator

A: aperture transmission function

A′: projection of the aperture on the detector

$\mathbf{A}^F$: finely sampled aperture array

$\mathbf{A}^{F\delta}$: finely sampled aperture array, δ decoding

G: decoding function

$\mathbf{G}^F$: finely sampled decoding array

$\mathbf{G}^{F\delta}$: finely sampled decoding array, δ decoding

$\mathbf{G}^0$: sampled decoding array resulting in no sidelobes

H: rect function

N: noise distribution

O: object intensity distribution

O′: pinhole image of an object

$\hat{\mathbf{O}}$: object estimate

R: recorded data

ψ: concentration parameter

a: object-to-mask distance

b: mask-to-detector distance

c_r: quadratic residues modulo r.

d_d: detector size

d_m: mask size

e: NTHT array parameter

f: illumination fractions

FoM: figure of merit

FoV: field of view (fully-coded)

m: magnification coefficient (coded aperture camera)

m_p: magnification coefficient (pinhole camera)

n: number of mask pixels (side); integer

p_d: detector pixel size

p_m: mask pixel size

r: radius

$\vec{r}$: position vector

t: mask transmission

t_m: thickness of one mask layer

x, y: coordinates

w_m: pinhole width

M: sidelobe height

N: total number of open holes in a mask

N_T: total number of positions in a mask

p_m: mask pixel size

S_m: shape of mask pixels

S_p: shape of detector pixels

S'_m: shape of the projection of a mask hole on the detector

z: object-to-mask distance

α: sampling parameter; incidence angle

δ: Dirac delta function

$\delta(i,j)$: Dirac delta function

$\Delta\vartheta$: angular field of view

θ: incidence angle

λ_g: geometric resolution

λ_s: system resolution

μ: attenuation coefficient; mean value

ρ: array open fraction (density)

σ: standard deviation

ξ: noise parameter

χ: parameter for the calculation of the effects of the intrinsic PSF on resolution

Ω: solid angle

Also see section 4.2 for a description of the symbols involved in SNR calculations.

LIST OF ACRONYMS

CAFNA: Coded Aperture Fast Neutron Analysis
FCFV: Fully Coded Field of View
FFT: Fast Fourier Transform
FNA: Fast Neutron Analysis
FWHM: Full Width at Half Maximum
FZP: Fresnel Zone Plate
MURA: Modified Uniformly Redundant Array
NRA: Non Redundant Array
NTHT: No-Two-Holes-Touching
PCFV: Partially Coded Field of View
PSF: Point Spread Function
SNR: Signal-to-Noise Ratio
URA: Uniformly Redundant Arrays

INTRODUCTION

The original motivation for our research group to study coded apertures was the development of a non-intrusive bulk inspection system, CAFNA (Coded Aperture Fast Neutron Analysis), which has been under development at MIT for a number of years. CAFNA is the combination of coded aperture imaging with Fast Neutron Analysis (FNA), an established bulk analysis technique based on the detection of γ-rays generated by inelastic scattering of fast (>1 MeV) neutrons. Since the energy of the photons emitted is specific to the isotopes present in the inspection volume, FNA is sensitive to the isotopic composition of materials. Information on the relative and absolute amount of common elements such as carbon, oxygen and nitrogen can uniquely identify a number of materials ([1], [2]), hence the interest in the technique for security applications (explosive detection in luggage or cargoes) and contraband detection. Since the cross sections and solid angles involved in FNA of a volume as large as a cargo container typically lead to poor statistics, it is imperative to make optimal use of the photons obtained. The idea at behind CAFNA is to form an image of the spatial distribution of the γ-rays generated in FNA by using a coded aperture in place of inefficient image-forming devices like collimators and pinholes.

The design of a CAFNA system requires work on two system components: a position and energy-sensitive γ-ray detector and the coded aperture. Of course the two problems are interdependent, the design of the detector depending on the design of the coded aperture and vice-versa. Both detector and aperture need significant development, but, while the considerations involved in the design of the detector can rely on a well-established body of knowledge, the design strategy of the coded aperture was largely uncharted territory to us. To get an approximate idea of the requirements on the detector we needed to tackle this latter problem first.

The complexity and size of a system such as CAFNA not only demanded theoretical investigation, but also verification by simulation and experiment, which requires use of a γ-ray detector. To this end, we needed to concentrate momentarily on a problem for which a fully-developed γ-ray position-sensitive detector is already available. This is the case of the Anger (or gamma) camera in Nuclear Medicine, so we started looking into the problem of taking a γ-ray picture of a thyroid (a classic test object in the field) with maximum resolution and Signal-to-Noise Ratio (SNR). At the beginning, the study focused on the determination of basic imaging parameters of the coded aperture camera (the combination of a coded aperture with a detector) such as field of view and resolution. Simulations

confirmed the predictions of the theoretical analysis, but also showed that other factors, in particular the object-to-detector distance, play an important role in the imaging process. This was not surprising, because we were aware that coded aperture imaging was originally developed for a very different application, X-ray space telescopes. The geometry of this case (a far-field problem, because the object can be considered very far from both coded aperture and detector) is radically different from that of our case (a near-field problem, because, for counting efficiency reasons, the object must be kept as close as possible to coded aperture and detector). In particular, all the apertures that in far-field ensure artifact-free imaging deviate significantly from ideality when used in near-field. The goal of devising a rational design procedure for the coded aperture was now expanding to include near-field artifact reduction. A second finding of the simulations was that, among the many apertures that provide ideal far-field imaging, some families have a SNR significantly superior to others, which opened yet another question, that of finding the optimal aperture family. This thesis describes the theoretical analysis developed to address these problems, the supporting simulations and experimental confirmation of the theory, and the solutions implemented in the design of a prototype coded aperture to answer all challenges in a rational and practical way.

There are two main results of this work. The first is the experimental demonstration of the feasibility of near-field artifact-free imaging over small fields of view. The high SNR achieved despite the high resolution (about 3 times better than achieved by state-of-the-art collimator and pinhole systems) is very promising for immediate application in routine small-animal laboratory planar studies. The design ideas presented in this thesis allows one to continue the development of a coded aperture camera for Nuclear Medicine along several lines. First, preliminary studies show that sub-millimeter resolution can be reached while retaining acceptable SNR. Even more interesting is the case of isotopes emitting high energy γ-rays, which, penetrating collimator septa, are very difficult to image with conventional methods. Second, the experience gained in the Nuclear Medicine study is the best validation of ideas and procedures general enough to be applied to CAFNA. A better understanding of the performance of coded apertures has allowed us to estimate the requirements of a satisfactory γ-ray detector. In particular, a prototype detector under development at MIT already seems to provide sufficient energy and position resolution. Its limit was rather found in its size, currently limited to an 8×8 array of 10×10 cm detectors, but can be easily overcome with translations in this experimental phase and by simply building a larger array in a full-scale system.

The results of the Nuclear Medicine application have met considerable interest in the medical arena, for both animal and human studies. The development of new pharmaceuticals typically passes through small animal studies whose aim is to identify if the compound is actually metabolized as

expected. Given the dimensions of the animals, resolution is key in these studies. With the 6-mm resolution characteristic of current methods it is very hard to identify parts of organs often smaller than a centimeter. Sub-millimeter resolution would provide researchers with a tool more practical than autoradiography techniques, which require painstaking surgical procedures and sacrifice of the animal, precluding time-dependent studies. Resolution is also of great interest in pediatric imaging, where reduced dimensions pose a challenge to current instrumentation, and in adult studies, such as breast imaging. This application is particularly suited to coded aperture imaging because it requires that a hot spot be located on a colder background. In this case better resolution can lead to all the benefits of an earlier diagnosis.

Given this potential, the investigation of coded apertures for nuclear medicine has become an independent project which will be hopefully continued in the future.

PART I:

INTRODUCTION AND BACKGROUND

Chapter 1 OVERVIEW AND HISTORICAL BACKGROUND

The goal of this Chapter is to put this thesis in context and introduce some problems of coded aperture imaging, so that the thesis outline can be discussed in some detail. Accordingly, only a broad overview is provided of questions that are going to be discussed in the next Chapters, to which rigor is postponed for clarity and brevity. Reviews of coded aperture imaging basics can be found in ref. [3]-[6].

1.1 The classic methods of 2d imaging

An image is a mapping over space of some distribution, in our case that of a photon emitter. At the energies of our interest (140 keV – 10.8 MeV)[1] wavelengths are small (8.9 pm – 115 fm), so that diffraction can be neglected and geometric optics used with excellent approximation. Placing a position sensitive detector directly before the emitting object (source) is not enough to generate an image because any photon detected (event) could be due to any part of the source (Figure 1.1a). In this sense, no spatial information is obtained and it is impossible to produce an image. An imaging system must associate

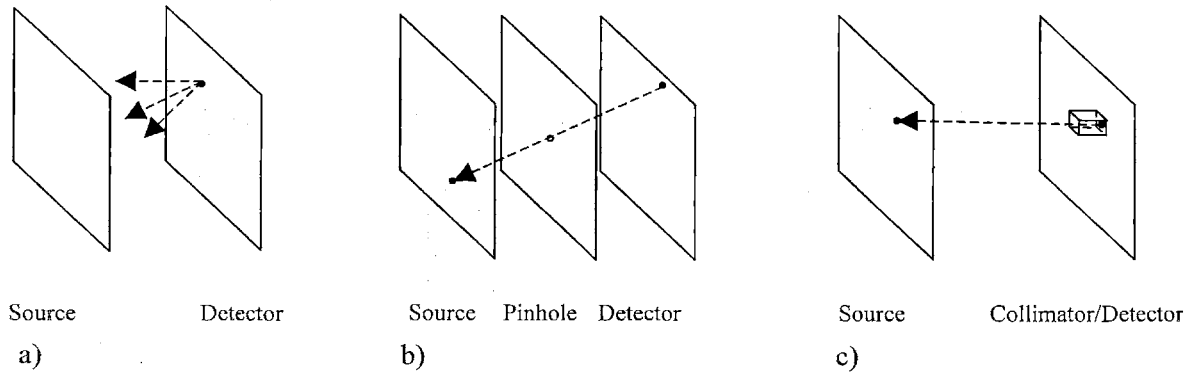


Figure 1.1: a) if no imaging system is present, a count collected at the detector can not be traced back to any specific part of the source. A pinhole (b) and a collimator (c) establish a one-to-one correspondence between detector and object. With a geometrical construction one can associate each detected event with an emission location.

[1] 140 keV is the energy of the γ-rays from the de-excitation of ^{99m}Tc, used in the great majority of Nuclear Medicine exams. 10.8 MeV is the energy of a thermal capture event on ^{14}N, a reaction used in explosive detection systems based on thermal neutrons.

events with a place of emission.

The simplest imaging device is the "pinhole", a slab of material opaque to radiation in which is poked an infinitely small (ideally dimensionless) hole. Since every event must have come from the object through the pinhole, along a straight line (Figure 1.1b), every point of the detector represents a point of the source and an image is formed. By inspection, the photon distribution recorded at the detector is an inverted picture of the object. A second system is the parallel-hole collimator, an array of infinitely small little tubes, whose walls are opaque to radiation, typically of hexagonal or circular section, placed side by side until the detector is covered. In this case, photons must have come from a line perpendicular to the detector (Figure 1.1c), which identifies the place of emission. The image is directly the photon distribution recorded by the detector.

Ideal pinholes and collimators share the property of realizing a one-to-one correspondence between object and image. Also, in both cases, for a given source, only photons traveling in one direction are collected. Lens systems also work with the same idea of one-to-one correspondence but, unlike pinholes and collimators, bend the path of incoming photons. This means that a point of the source can

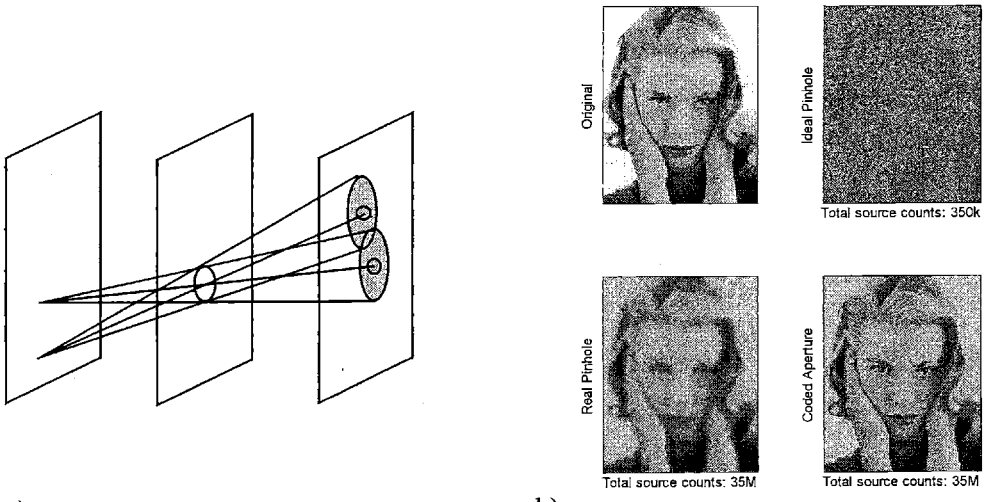


Figure 1.2: a) resolution loss when an ideal pinhole is enlarged to increase throughput. b) intuitive visualization of the trade-off between noise and resolution for pinhole and collimator imagers. Three pictures of the original object (top left) were simulated for constant exposure time. The ideal pinhole was 1×1 image-pixels wide (1:1 magnification). The image is very noisy (top right). To obtain better statistics one can widen the pinhole to 10×10 image pixels and collect 100 times more counts. This does improve the signal in the image (bottom left) but also blurs it. In a half-open 15×15 coded aperture there are approximately 100 1×1 pinholes: hopefully, this yields the same signal advantage of the larger pinhole while preserving resolution (bottom right). In Chapter 4 we will see that this argument holds for a point-like source only. For images as complex as the one chosen here, depending on its statistical details, there may be no SNR advantage in using a coded aperture.

contribute a whole cone of different directions to the image, with great advantage for the SNR. Unfortunately, photons of the energies of our interest can not be bent by refraction optics. Bragg diffraction mirrors used in space telescopes work well up to 15 keV ([3]), but further extension requires sophisticated manufacturing techniques and does not exceed 80 keV.

In theory, the resolution provided by an ideal pinhole (i.e. a dimensionless point) is perfect. An intuitive argument is that two point sources arbitrarily close in the object will always show separated on the detector[2]. However, this comes at the price of no counts at all, because the area of the ideal pinhole is zero and the photon flux through it must be zero as well. This is also true of collimators, but not of lenses, which offer a finite area to the incoming flux, but, again, do not work at the energies of our interest. Real pinholes, however, must have a finite size. This allows some photons to pass, which does increase the SNR, but does not come completely to our rescue. Figure 1.2a shows that if two point sources are close enough, their projections on the detector, which would be distinct in the case of an ideal pinhole, are not separated. Resolution must have decreased. Another way of looking at the same issue is to recognize that the larger pinhole realizes only an "approximate" one-to-one correspondence. For a complex object, resolution loss means a blurred image. This is shown pictorially in the example of Figure 1.2b, where the ideal pinhole (in this example a very small hole of finite size) is compared to a real one.

In conclusion, the pinhole (or collimator) hole size can not be increased indefinitely to increase efficiency because some resolution limit will be reached. In typical collimator systems only 0.1% or less of the emitted photons are counted, giving a noisy image unless a long exposure time is used or lower resolution accepted for a constant exposure time.

1.2 Why a coded aperture?

Coded apertures try to achieve the resolution of small pinholes while maintaining a high signal throughput. The basic idea is to overcome photon shortage by opening many small pinholes instead of a larger one. These pinholes are placed in specially designed arrays called patterns. The aperture (or mask) is the physical realization of a pattern. The mask forms with the detector the coded aperture camera.

---

[2] Since here we are concerned with the properties of the imaging optics, not with the system as a whole, an ideal (perfect resolution) detector is assumed. A complete discussion of the resolution of a coded aperture camera is given in sections 2.6 and 7.1.

1.3 Coded aperture imaging in a nutshell

Since the number of photons passing through a pinhole of the coded aperture is independent of photons passing through all other pinholes, each pinhole is independent of the others. The projection of the object through the mask can be decomposed in the sum of contributions from each pinhole. From the discussion of section 1.1, a pinhole casts on the detector an inverted image of the object, which superimposes to projections from other pinholes (Figure 1.3a). The counts collected at the detector are, then, the superposition of many shifted copies of the object. In a far-field approximation, i.e. when the object is sufficiently far from mask and detector, the projection process follows the equation:

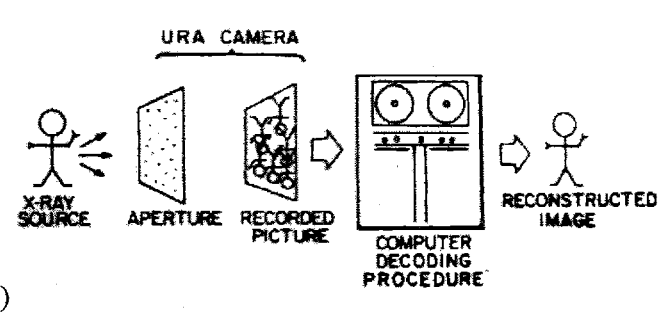
$$\mathbf{O} \times \mathbf{A} = \mathbf{R} \tag{1.1}$$

where **O** is the irradiance (number of photons emitted per unit area) of the object, **A** the transmission of the coded aperture (a function ranging from 0 for complete opacity to 1 for complete transparency), **R** the counts recorded by the detector and × indicates non-periodic correlation. A rigorous definition of the far-field approximation is given in Chapter 5, where eq. (1.1) is derived as a particular case of a more general formulation.

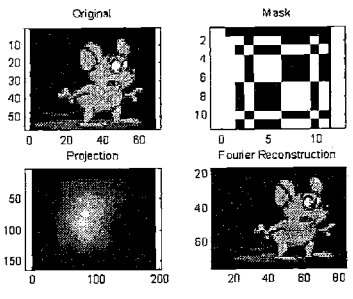
As the pinholes can be several hundreds, **R** does not resemble **O** in any immediate way. An alternative way of looking at the same process is to say that a point in the image is not represented on the detector by a point, but rather by a pattern of points. This is the mask itself, as follows from eq. (1.1), when **O** is replaced with Dirac's delta function δ:

$$\mathbf{R} = \delta \times \mathbf{A} = \mathbf{A} \tag{1.2}$$

Therefore, each point source is present in the projection not as a point but as a known pattern.



a)



b)

Figure 1.3: a) pictorial summary of a coded aperture camera concept (here indicated with a URA camera). Adapted from [6]. b) a sample of the process from the object, through the mask, to projection and reconstruction.

Different point sources are characterized by the pattern shift. In this sense the signal from the source is encoded. The first consequence is that a point source is not counted once, but once for every pinhole of the coded aperture, which is expected to increase the counting statistics and, consequently, the SNR. The second consequence is that at each detector point is present information about many points of the source: in this sense information is multiplexed. The third consequence is that the recorded pattern **R** must be decoded to obtain an image. In intuitive terms, **R** must be scanned looking for known patterns which must be replaced with the point source that cast them. This is the same as separating the overlapped copies and is done by an operation known in signal processing as "matched filtering" ([5]). The technique prescribes to take the correlation of the collected data with the known pattern, in our case **A**. In a more general case in a pattern may be sought not through the pattern itself, but through an associated decoding pattern (or decoding array) **G** such that:

$$\mathbf{A} \otimes \mathbf{G} = \delta \tag{1.3}$$

where $\otimes$ indicates periodic correlation, the matched filtering process is:

$$\mathbf{R} \otimes \mathbf{G} \tag{1.4}$$

The result of this operation is to produce a perfect copy of the object **O**. In fact, given the linearity of correlation operations and eq. (1.1) one can write (using Appendix A.3):

$$\hat{\mathbf{O}} \equiv \mathbf{R} \otimes \mathbf{G} = (\mathbf{O} \times \mathbf{A}) \otimes \mathbf{G} = \mathbf{O} * (\mathbf{A} \otimes \mathbf{G}) \tag{1.5}$$

where $\hat{\mathbf{O}}$ is by definition the estimate of the object or reconstructed image. This chain of equalities shows that the output of the imaging system is not directly the object but, as in all linear systems, a convolution of the object with a kernel, in this case $\mathbf{A} \otimes \mathbf{G}$. The convolution kernel is also called the Point Spread Function (PSF), which is the imaging analogue of the Pulse Response Function of electrical circuits. With this definition, eq. (1.5) becomes:

$$\hat{\mathbf{O}} = \mathbf{O} * PSF \tag{1.6}$$

The name Point Spread Function comes from the fact that the PSF is the image produced in response to a point source. In fact if $\mathbf{O} = \delta$:

$$\hat{\mathbf{O}} = \delta * PSF = PSF \tag{1.7}$$

which means that the PSF describes the imperfections that cause a system not to reconstruct a point with a point, but to spread it over a certain area. The PSF summarizes the behavior of the imaging system because the output of the imager can be predicted from knowledge of the input (the object) and the PSF alone, via eq. (1.6). Its importance is enormous both in theory and in practice. From eq. (1.5) and (1.6), in coded aperture imaging:

$$PSF = \mathbf{A} \otimes \mathbf{G} \tag{1.8}$$

Fortunately **A** and **G** are both in the hands of the designer. Furthermore, considerable literature is dedicated to the generation of pairs of **A** and **G** satisfying the constraint of eq. (1.3). Such pairs are said to have perfect imaging properties. In fact, substitution of eq. (1.3) in eq. (1.8) and then in eq. (1.6) gives:

$$\hat{\mathbf{O}} = \mathbf{O} * \delta = \mathbf{O} \tag{1.9}$$

which means that if the *PSF* is a δ function the reconstruction is perfect. This should not be surprising, because in this case a point in the object corresponds to a point, and not a blur, in the image.

In conclusion, coded aperture imaging can produce a perfect copy of the object. It is a two step process: the first is physical, the projection of the source through the aperture; the second is computational, decoding. The motivation to go through this complication is the potential of achieving a higher SNR. Chapter 4 is dedicated to quantifying this potential and understanding the hypothesis under which it is actually present.

Of course, the result of ideal reconstruction is due to a number of hypotheses. The first, and most relevant, is that eq. (1.1) holds only in the above-mentioned far-field approximation, which is the implicit starting point of all coded aperture literature. This approximation does hold in most literature applications, especially the early ones, which developed the basic ideas of coded aperture imaging in the context of space applications. However, in Nuclear Medicine, the concern is to collect the maximum number of photons, and the detector must be placed as close as possible to the source. In typical cases, the far-field approximation breaks down, but little attention to this is found in published works. Chapter 5 is dedicated to the examination of the consequences of using methods developed for far-field applications in a near-field geometry and the development of suitable remedies.

The second major hypothesis is that detector and mask be ideal. As for the former, a typical state-of-the-art Anger camera provides a 3.7-mm-FWHM (Full Width at Half Maximum) PSF at the center of the crystal. This means that an infinitely narrow beam is seen not as a point but as a blur reaching half of its peak value only outside a diameter of 3.7 mm. This figure gives a first idea of the resolution that can

be reached with such a system, but it has to be combined with a second factor. In fact, eq. (1.3) holds in this exact form only for an aperture with dimensionless pinholes. Real pinholes and collimators have a finite size, which further degrades resolution. As an example, the Ultra-High-Resolution collimator supplied by Siemens for use with its E-Cam gamma-camera has a hole diameter of 1.16 mm and is capable of a system resolution of 6.3 mm for ^{99m}Tc at 10 cm. This collimator, however, has the low sensitivity (counts per unit activity in the source) of 100 cpm / μCi at 10 cm. For comparison, a High-Sensitivity collimator (1063 cpm / μCi at 10 cm) has a resolution of 14.6 mm at 10 cm. In Chapter 7 is provided a thorough description of how all these factors were combined in the determination of the resolution of a coded aperture camera and how the limits for the achievable resolution were investigated.

1.4 Coded aperture history

Coded aperture techniques were first proposed in 1961 by Mertz and Young ([7]). The aperture they proposed was the Fresnel Zone Plate (FZP), in theory a circularly-symmetric mask having transmission:

$$\cos(r^2); \quad 0 \le r \le +\infty \tag{1.10}$$

where r is the radius from the center of symmetry. This pattern was inspired by holography, where it is used for its property of refocusing coherent light in a focal point. This characteristic can be used to decode the projection. In early experiments the projection was recorded on a film, which was developed and then exposed to coherent light of wavelength comparable to the size of the projected pattern. Since each projected zone plate is refocused in a point, the image is decoded. Note that while the projection is cast by X or γ-rays, coherent radiation of optical wavelength is used in decoding, which was an optical

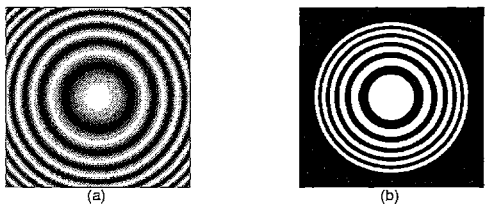


Figure 1.4: theoretical (a) and practical (b) Fresnel Zone Plate. The latter is non-ideal because transmission is either total or null (instead of being continuously modulated) and because the plate is not infinite but stops after a few circles.

procedure ([6]).

The advent of fast digital computers made it possible to exploit a second nice characteristic of the FZP: the auto-correlation of an FZP is a δ function ([3]), so the decoding method of section 1.3 can also be used. However, perfect imaging properties hold only if the plate is infinite and has a continuously varying transmission. Since this entails considerable fabrication difficulties, in real applications the FZP must be approximated with a finite series of concentric circles of radius:

$$r_n = r_0\sqrt{n}, \quad n = 1, 2, 3\ldots, n_{max} \tag{1.11}$$

where the annuli are alternatively totally opaque and transparent (Figure 1.4). A particular case is that of the annulus, where a single open ring is used. These approximations cause significant deviation from a δ function, even if more general formulations of eq. (1.11) are used ([8]). This, however, did not stop early experiments. Despite having been proposed for space applications, the first actual demonstration of FZP imaging was a study of a thyroid phantom ([9]). Fabrication difficulties, with the problems associated with optical decoding, made the technique impractical.

The advent of more manageable apertures gave new momentum to the field. In 1968, Dicke and Ables independently pointed out that a square arrangement of randomly distributed square openings (a random array) has reasonable self-correlation properties ([10], [11]). Unfortunately, just like the FZP, a random aperture provides an ideal PSF only if it is infinite. In 1971 the Non Redundant Arrays (NRAs, [12]) were proposed. These arrays are compact but have ideal properties only on a small field of view and contain a small number of holes, which prevents great improvements in the SNR. The difficulty was overcome in 1978, when Fenimore and Cannon introduced the rectangular Uniformly Redundant Arrays (URAs, [13]), which have an ideal PSF and are finite. A decade later URAs were followed by the Modified URAs (MURAs), which have the additional convenience of being square ([14]). Meanwhile, a number of other apertures were discovered. They are described in section 2.4.

1.5 Applications of coded apertures

Applications of coded apertures have been, for the vast majority, in astronomy. A number of examples can be found in ref. [3]. This is due to two reasons. The first has to do with the appearance of the object (in astronomy a number of isolated bright spots, the stars) and the detector high-background environment, often an orbiting satellite or a balloon-borne telescope. As we shall see, it is in these conditions that coded apertures provide the largest SNR advantages over pinholes and collimators. The

second reason is that star imaging is a perfect example of far-field imaging, a condition not affected by artifacts.

Other fields of application are nuclear medicine ([9], [15]-[18]), nuclear fusion ([19], [20]), industrial imaging, e.g. clean-up and decommissioning of nuclear sites ([21]), contraband detection ([22]), chemical spectroscopy and optical image processing ([3]).

1.6 Thesis outline

The thesis is divided in three parts. As the reader will have found out by now, Part I provides context and a general overview. The second Chapter is a more detailed presentation of coded aperture imaging. An analysis of the imaging geometry is presented with the goal of deriving fundamental relations among basic parameters such as field of view and resolution. For comparison reasons, the case of the pinhole is also briefly summarized. Extensive details on the options available for the aperture and decoding pattern are provided. Listing all pattern families and generation rules may seem tedious and not original, but the goal was to provide some order in information otherwise scattered in a number of papers. Furthermore, a reasonably comprehensive knowledge of patterns is very useful in understanding SNR and artifact reduction problems. Some considerations on the 3d properties of coded apertures close the Chapter.

Part II of the thesis is its original core. It discusses the simulation and theoretical tools used in the investigation, how they were used in the design and construction of a prototype aperture and the experimental results obtained.

Simulation has played a fundamental role in this project to the point that it is not unfair to say its results directed the work. Since some aspects of simulation are not trivial, Chapter 3 describes the simulation code. In this Chapter is also included a description of the optical simulator that provided the first link between the computational and the real world, supporting the credibility of computer calculations.

Chapters 4 and 5 concern, respectively, the SNR and artifact reduction. A major problem was finding a rational procedure to design the coded aperture. For example, from literature are known several families of pairs (**A**, **G**), and, within each family, patterns of many different sizes. Since all families and patterns provide ideal far-field imaging, the problem was to find some other criterion that made a family preferable to others. To answer this question, Chapter 4 looks closely at the SNR of coded aperture families. One of the conclusions is that coded apertures do not always provide an advantage over a pinhole or collimator system. In particular, a coded aperture is advantageous for sparse objects (a result

well-known in the literature), i.e. objects for which the activity is concentrated in a relatively small part (on the order of 10% or less) of the field of view, unless a very high background is present. A less-known result is that low-throughput apertures do not always provide better performance for sparse objects, a statement that has led some researchers to erroneously propose some aperture families as optimal for medical applications. Finally, investigations in the literature have considered the ideal system: for instance, the area around mask holes is perfectly opaque and far-field geometry is assumed. In the analysis of the SNR partial mask transparency is introduced because, while it does not introduce any artifacts, it does reduce the SNR. Results extend and correct some of those previously published and will be later applied to the determination of mask thickness and in the investigation of higher energy isotopes.

In Chapter 5 several other non-idealities are presented and their impact on the final image described. These are mainly related to how the detector samples the projected pattern, but the most interesting for planar imaging are those due to near-field geometry and mask thickness. The far-field approximation is derived as a particular case of a more general mathematical framework capable of predicting near-field artifacts, which are described, classified and compared to published results. On these bases, some remedies are proposed and their effectiveness tested with computational simulations. The Chapter closes with a description of the origin of thickness artifacts.

The pairs (**A**, **G**) are found in literature as dimensionless 2d arrays of numbers. The question of finding the physical dimensions to fabricate a pattern into a mask is closely related to that of finding the field of view and the resolution of a coded aperture camera. Chapter 6 tackles the problem in a systematic way by applying to the design of a coded aperture the concepts developed in previous chapters. The reason for every value of a mask specification is explained. The mask designed was actually built and experimental results are presented.

Part III contains advanced topics and extensions of the work presented in Part II. Chapter 7 revisits the resolution of the prototype mask to include the effect of detector sampling and intrinsic PSF. After different detector setups are discussed, the characteristics of an advanced mask design, capable of sub-millimeter resolution, are derived. Since the ultimate limit on resolution is found to be mask thickness, a method for an accurate determination of optimal thickness is developed.

Chapter 8 is about future work that can be done with the prototype mask, the bases for the imaging of higher-energy isotopes and the general issues involved in 3d coded aperture imaging.

Finally, Chapter 9 introduces fast neutron analysis and presents experimental results which provide insight on issues, such as background and noise, likely to play a prominent role in the design of a CAFNA system. The details of a candidate mask for CAFNA are also given.

Chapter 2 FUNDAMENTALS OF CODED APERTURE IMAGING

This Chapter explores the fundamentals of coded aperture imaging more in depth. First, the pinhole camera gives an opportunity to introduce at an intuitive level many ideas useful for what follows. Different aperture families, code aperture camera geometries and decoding strategies are then presented along with many different families of apertures. A few considerations on elementary three-dimensional properties of coded aperture imaging close the Chapter.

A basic exposition of coded aperture imaging that also provides historical background can be found in ref. [6]. Ref. [13] is more technical and is probably the most referenced paper in the field.

2.1 The pinhole camera

The photon distribution recorded at the detector of a pinhole camera follows from the one-to-one correspondence established between object and detector. With the definitions of Figure 2.1, the photon distribution at a generic detector position $R(x_i, y_i)$, must be due to the point source at (x_o, y_o) only, to whose irradiance $O(x_o, y_o)$ must be proportional:

$$R(x_i, y_i) \propto O(x_o, y_o) \tag{2.1}$$

Defining the vectors:

$$\vec{r}_i = (x_i, y_i) \text{ and } \vec{r}_o = (x_o, y_o) \tag{2.2}$$

which, because the ray going from $\vec{r}_o$ to $\vec{r}_i$ must pass through the pinhole, are related by:

$$\vec{r}_o = -\frac{a}{b}\vec{r}_i \tag{2.3}$$

eq. (2.1) becomes:

$$R(\vec{r}_i) \propto O\left(-\frac{a}{b}\vec{r}_i\right) \tag{2.4}$$

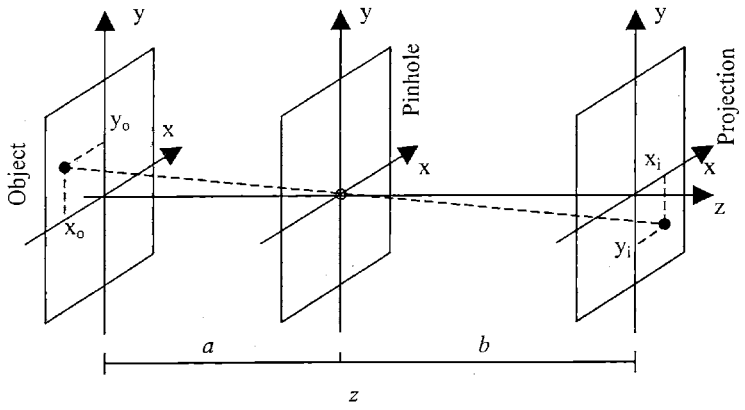


Figure 2.1: pinhole camera geometry.

This relationship shows that the projection through the pinhole is a copy of the object, inverted (because of the minus) and rescaled. Since a and b can be any positive number, the rescaling constant can be any positive number. If $a > b$, i.e. the pinhole is closer to the detector than to the object, the object appears minified, while for $a < b$, the object is magnified. From Figure 2.2a one can verify that the ratio of the projected object size to the original is (neglecting the inversion):

$$\frac{h_i}{h_o} = \frac{b}{a} = m_p \tag{2.5}$$

which demonstrates that the rescaling coefficient is the magnification coefficient of the pinhole m_p.

If h_i is set equal to the size of the detector, d_d, the size of the field of view of the pinhole, i.e. the set of points in the plane of the object that can be imaged, is obtained:

$$FoV = \frac{d_d}{m_p} \tag{2.6}$$

In the case of the ideal pinhole, resolution is, of course, perfect. If the pinhole had a finite width w_m, each point would cast an image of size:

$$w_d = \frac{a+b}{a} w_m = \left(1 + m_p\right) w_m \tag{2.7}$$

If the resolution of the system is defined as the minimum distance between two points in the plane of the object such that their projections are separated in the image, from Figure 2.2b and simple geometry:

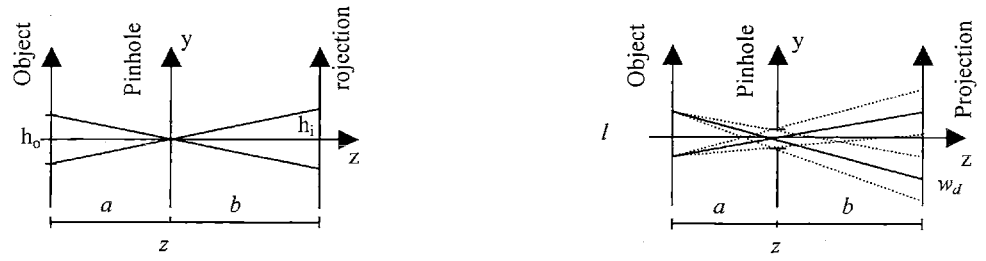


Figure 2.2: a) Determination of the pinhole magnification coefficient b) Determination of pinhole resolution.

$$l \geq \frac{a}{b} w_d = \left(1 + \frac{1}{m_p}\right) w_m \tag{2.8}$$

With this definition of resolution, low values indicate good resolution. This clarified, this equation shows that the best case is that of infinite magnification and that resolution is limited by the size of the pinhole. It is interesting to note that while resolution improves for increasing magnification, the field of view shrinks. The ratio of the two is:

$$\frac{FoV}{l} = \frac{d_d}{\left(1 + m_p\right) w_m} \tag{2.9}$$

This ratio could be taken as a figure of merit for an imager, which ideally should have the widest possible field of view and, at the same time, the best possible resolution. The maximum value is d_d / w_m and is obtained for $m_p \to \infty$. As magnification increases, resolution improves but eventually reaches the asymptote w_m while the field of view decreases indefinitely until the figure of merit vanishes.

2.2 Encoding the signal: object projection

Chapter 1 showed that coded aperture imaging is a two step process. The first step, encoding, is the physical process of projection of the object, through the mask, onto the detector. Since at the energies of interest geometric optics is applicable, it is possible to calculate the projection from aperture and object with a purely geometrical argument. With the definitions of Chapter 1 and Figure 2.3a, the photon distribution R recorded at the detector position $\vec{r}_i$ and due to the point source at $\vec{r}_o$ must be proportional

to the irradiance $O(\vec{r}_o)$, modulated by the transmission of the mask A evaluated at the point of intersection with the ray going from $\vec{r}_o$ to $\vec{r}_i$:

$$R(\vec{r}_i) \propto O(\vec{r}_o) A\left(\vec{r}_o + \frac{\vec{r}_i - \vec{r}_o}{z} a\right) \tag{2.10}$$

To simplify construction, A is almost always considered a two-valued function. It is also very often thought as a grid of square elements, so that it can be represented with a matrix whose elements are 1s and 0s (respectively, holes and opaque elements). To obtain the total recorded photon distribution, it is sufficient to repeat this argument for all point sources and add all the results, i.e. integrate over the object plane:

$$R(\vec{r}_i) \propto \iint_{\vec{r}_o} O(\vec{r}_o) A\left(\vec{r}_o + \frac{\vec{r}_i - \vec{r}_o}{z} a\right) d^2\vec{r}_o \tag{2.11}$$

Here the assumption that the detector responds linearly is made, i.e. the response to the simultaneous exposition to several sources is equal to the sum of the responses obtained from the sources taken one at a time. By definition:

$$z = a + b \tag{2.12}$$

so that:

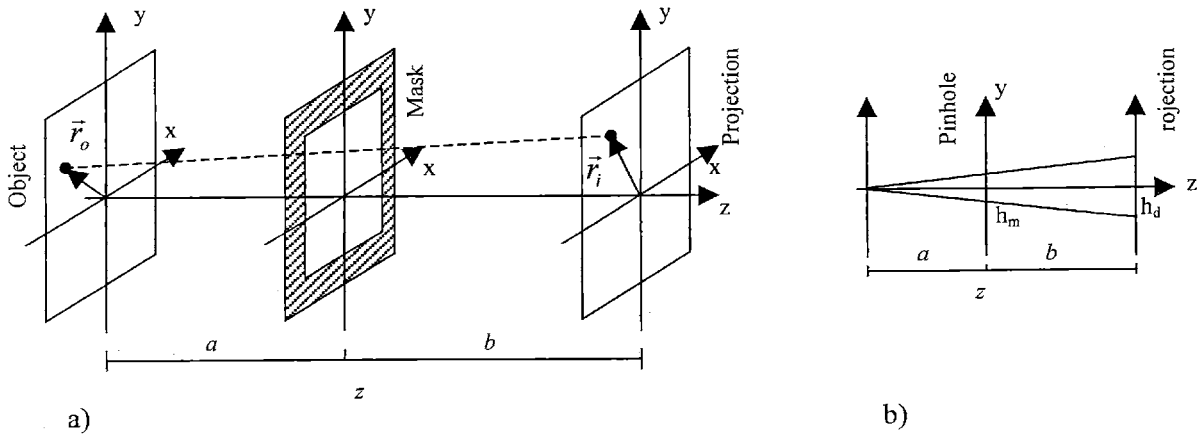


Figure 2.3: a) projection geometry. b) calculation of the magnification coefficient of the mask

$$R(\vec{r}_i) \propto \iint_{\vec{r}_o} O(\vec{r}_o) A\left(\frac{a\vec{r}_i + b\vec{r}_o}{z}\right) d^2\vec{r}_o \tag{2.13}$$

which, with:

$$\vec{r}_o^{\,i} = -\frac{b}{a}\vec{r}_o \tag{2.14}$$

becomes:

$$R(\vec{r}_i) \propto \iint_{\vec{r}_o^{\,i}} O\left(-\frac{a}{b}\vec{r}_o^{\,i}\right) A\left(\frac{a}{z}\left(\vec{r}_i - \vec{r}_o^{\,i}\right)\right) d^2\vec{r}_o^{\,i} \tag{2.15}$$

From its definition, $\vec{r}_o^{\,i}$ is the point that would be associated to $\vec{r}_o$ in a pinhole imager, i.e. it is the point of the detector aligned with $\vec{r}_o$ and the mask center. Two more definitions will help cast this equation in a more meaningful form:

$$O'(\vec{r}) \equiv O\left(-\frac{a}{b}\vec{r}\right) \text{ and } A'(\vec{r}) \equiv A\left(\frac{a}{z}\vec{r}\right) \tag{2.16a-b}$$

O' is a scaled and inverted version of O. The scaling coefficient is the same as the magnification coefficient of a pinhole camera (see eq. (2.5)). In other words, O' is a pinhole image of the object (eq. (2.4)). Similarly, A' is a scaled (but not inverted) version of A. The scaling coefficient makes A' larger than A. Figure 2.3b helps in the calculation of the ratio of the size of the mask projection to the size of the mask itself:

$$\frac{h_d}{h_m} = \frac{z}{a} \tag{2.17}$$

which is the scaling coefficient of definition 2.16b. The magnification of A' is due to the projection of the mask pattern on the detector. Substitution of eq. (2.16a-b) in eq. (2.15) gives:

$$R(\vec{r}_i) \propto \iint_{\vec{r}_o^{\,i}} O'(\vec{r}_o^{\,i}) A'(\vec{r}_i - \vec{r}_o^{\,i}) d^2\vec{r}_o^{\,i} = O' * A' \tag{2.18}$$

which shows that the projection process is described by the convolution[3] of the pinhole image of the object O' with the projection of the mask pattern A'. A physical interpretation of this equation is that the projection is the sum of magnified mask patterns, each shifted according to the location $\vec{r}_o$ of the point source casting the shadow and weighted according to its irradiance.

An alternative interpretation is obtained exploiting the commutative property of convolution:

$$R = O' * A' = A' * O' = \iint_{\vec{r}_m^i} A'\left(\vec{r}_m^i\right) O'\left(\vec{r}_i - \vec{r}_m^i\right) d^2\vec{r}_m^i \tag{2.19}$$

where $\vec{r}_m^i$ is a dummy variable spanning the detector. In this form the projection is the sum of equal pinhole images of the object, shifted and weighted according to various positions in the mask. From Chapter 1 we know that the aperture is in practice a collection of N pinholes. Assuming ideal pinholes, A can be written as a sum of shifted δ functions:

$$A(\vec{r}_m) = \sum_{n=1,2,...N} \delta\left(\vec{r}_m - \vec{r}_{m,n}\right) \tag{2.20}$$

where $\vec{r}_m$ is a generic position on the mask plane and $\vec{r}_{m,n}$ the position of the nth pinhole on the plane. The projection of A on the detector is:

$$A'(\vec{r}_m^i) = \sum_{n=1,2,...N} \delta\left(\vec{r}_m^i - \vec{r}_{m,n}^i\right) \tag{2.21}$$

which can be substituted in eq. (2.19) to get:

$$R = \iint_{\vec{r}_m^i} \sum_{n=1,2,...N} \delta\left(\vec{r}_m^i - \vec{r}_{m,n}^i\right) O'\left(\vec{r}_i - \vec{r}_m^i\right) d^2\vec{r}_m^i = \sum_{n=1,2,...N} O'\left(\vec{r}_i - \vec{r}_{m,n}^i\right) \tag{2.22}$$

---

[3] In several papers this convolution is reported to be a correlation. This difference comes from a different definitions of terms or a different setting of the axes on the object or the detector. For instance, for simplicity, in Chapter 1 the object was not inverted. This is equivalent to defining $\vec{r}_o^i = \frac{b}{a}\vec{r}_o$, which means an inversion of one set of axes. This definition leads to: $R\left(\vec{r}_i\right) \propto \iint_{\vec{r}_o^i} O\left(\frac{a}{b}\vec{r}_o^i\right) A\left(\frac{a}{z}\left(\vec{r}_i + \vec{r}_o^i\right)\right) d\vec{r}_o^i$, which is the correlation of eq. (1.1).

To establish the connection between $\vec{r}_m^{\,i}$ and $\vec{r}_m$ we can simply recall that the former is the projection of the latter on the detector, so, from Figure 2.3b:

$$\vec{r}_m^{\,i} = \frac{z}{a}\vec{r}_m \tag{2.23}$$[4]

In conclusion:

$$R = \sum_{n=1,2,\ldots N} O'\left(\vec{r}_i - \frac{z}{a}\vec{r}_{m,n}\right) \tag{2.24}$$

In this form the projection is the sum of N shifted copies of the pinhole image of object. While this point of view is useful in establishing a relationship with the pinhole camera, the decomposition of the projection into the sum of mask patterns is useful in introducing the decoding step. As we shall see in the next section, the procedure of decoding can be thought of as one of scanning the projection to identify mask projections.

2.3 Decoding: computer post-processing

The second step of coded aperture imaging is extracting the encoded data, i.e. the object from the convolution of eq. (2.18):

$$R(\vec{r}_i) \propto O' * A' \tag{2.18}$$

The classic strategy makes use of the convolution theorem:

$$\mathcal{F}(R) \propto \mathcal{F}(O') \cdot \mathcal{F}(A') \tag{2.25}$$

and the inverse Fourier transform to obtain the object estimate $\hat{O}$:

$$\hat{O} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(R)}{\mathcal{F}(A')}\right) = O' \tag{2.26}$$

---

[4] Also note that with this and definition 2.16b: $A'(\vec{r}_m^{\,i}) = A\left(\frac{a}{z}\vec{r}_m^{\,i}\right) = A(\vec{r}_m)$

Unfortunately this works only in absence of noise. In a real case, in fact, some noise N adds to the recorded data:

$$R(\vec{r}_i) \propto O' * A' + N \tag{2.27}$$

Fourier-transformation of both sides and rearrangement yields:

$$\hat{O} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(R)}{\mathcal{F}(A')}\right) - \mathcal{F}^{-1}\left(\frac{\mathcal{F}(N)}{\mathcal{F}(A')}\right) = O' - \mathcal{F}^{-1}\left(\frac{\mathcal{F}(N)}{\mathcal{F}(A')}\right) \tag{2.28}$$

In the assumption of white noise, $\mathcal{F}(N)$ is constant at all frequencies. On the other hand, $\mathcal{F}(A')$ typically has zeros. The net result is that the reconstruction of O' is dominated by the noise term. While improvements can be obtained by Wiener filtering ([23]), a different strategy gives better results. In section 2.2 was shown that each point of the source is present in the projection because it deposits a shadow of the mask on the detector. The correlation method of decoding is a way of locating mask patterns in the projection. The reconstructed object is defined to be:

$$\hat{O} \equiv R \times G \tag{2.29}$$

Substituting the expression for R:

$$\hat{O} = (O' * A' + N) \times G = (O' * A') \times G + N \times G = O' * (A' \times G) + N \times G \tag{2.30}$$

But if:

$$A' \times G = \delta \tag{2.31}$$

eq. (2.30) becomes:

$$\hat{O} = O' + N \times G \tag{2.32}$$

The noise term is still present but, unlike in the Fourier transform method, is not ill-behaved. On the contrary, $N \times G$ is the convolution of a constant (only in an average sense and in the assumption of uniform noise) with a function, which is a constant no matter what the second function may be (see A.2).

To gain some insight, it is convenient to neglect noise, which, anyway, contributes a constant background to the image. So, let R be:

$$R(\vec{r}_i) \propto O' * A' = \iint_{\vec{r}_o^i} O'(\vec{r}_o^i) A'(\vec{r}_i - \vec{r}_o^i) d^2\vec{r}_o^i \tag{2.33}$$

From the definition, the reconstructed image is, inverting the integration order:

$$\hat{O} \equiv R \times G = \iint_{\vec{r}_i} R(\vec{r}_i) G(\vec{r}_i + \vec{r}_r) d^2\vec{r}_i = \iint_{\vec{r}_o^i} O'(\vec{r}_o^i) \iint_{\vec{r}_i} A'(\vec{r}_i - \vec{r}_o^i) G(\vec{r}_i + \vec{r}_r) d^2\vec{r}_i d^2\vec{r}_o^i \tag{2.34}$$

What a correlation does is to take the decoding pattern, G, shift it (by adding the reconstruction variable $\vec{r}_r$ to the argument $\vec{r}_i$), multiply it point by point with the data R and then add all products (integration). The result is the brightness of the image at the reconstruction point $\vec{r}_r$. The procedure is then repeated for all shifts ($\vec{r}_r$) to complete the image. Since the operation is linear, we were able to invert the integration order: linearity reduces the problem to the simpler one of correlating G with a shape independent of the object, the mask A, or, better, its projection A'. G is a decoding array stored in the computer. A common choice is that it be also made of 0s and 1s, just as, in most cases, A. Indeed, it is not uncommon that $G = A$. In this case, when the reconstruction shift (the shift of G) with respect to A is 0, the 1s of G and A coincide. Point by point multiplication and successive summation gives the number of holes, which is the maximum possible result. For all other shifts some of the holes (1s) of A fall on the 0s of G and do not contribute to the sum. Importantly, if the result were the same for all these other shifts, the image would be a bright spot surrounded by a constant background (that can be set to zero by subtraction). The condition $A \times G = \delta$, therefore, can be interpreted as follows: the correlation tries all possible shifts and indicates the one matching the shift of the mask projection with a bright spot. Since this shift depends on the position of the source, the bright spot indicates the position of the source. Mathematically:

$$\iint_{\vec{r}_i} A'(\vec{r}_i - \vec{r}_o^i) G(\vec{r}_i + \vec{r}_r) d^2\vec{r}_i = \delta(\vec{r}_o^i + \vec{r}_r) \tag{2.35}$$

The δ function sifts the object O', selecting one point source at a time. In fact:

$$\hat{O} = \iint_{\vec{r}_o^i} O'(\vec{r}_o^i) \iint_{\vec{r}_i} A'(\vec{r}_i - \vec{r}_o^i) G(\vec{r}_i + \vec{r}_r) d^2\vec{r}_i d^2\vec{r}_o^i = \iint_{\vec{r}_o^i} O'(\vec{r}_o^i) \delta(\vec{r}_o^i + \vec{r}_r) d^2\vec{r}_o^i = O'(-\vec{r}_r) = \Re(O') \tag{2.36}$$

where $\Re$ is the reflection operator, which restores the reflection in the definition of O'.

In conclusion, reconstruction is the identification of a known pattern in a signal. This process is known in communication theory as "matched filtering", which is the optimal strategy for detecting a known signal in noise when the SNR tends to 0 ([5]). The drawback is that the design of the coded aperture is restricted to patterns having a very particular property. Fortunately, such patterns are not uncommon and different families of such patterns are available in literature. They are described in the next section.

2.4 Coded aperture families

In this section are described some families of coded aperture arrays, their generation rules and correlation properties. Even though a library of arrays would be of great help to coded aperture designers, we are not aware of any complete listing. The following is a reasonably comprehensive collection of arrays found scattered in many papers and reviews, but is very far from complete.

Literature nomenclature is very confusing. There are several instances of the same name being used for different sets or for a set and one of its subsets. For instance Caroli et al. ([3]) use the term Pseudo-Noise (PN) sequence as an equivalent of *m*- sequence, Calabro and Wolf use it for twin-prime Hadamard sets ([25]), while Nelson and Fredman use it for all Hadamard sequences ([26]), which were later recognized to be Uniformly Redundant Arrays ([27]), a term initially reserved to Hadamard twin-prime sequences ([13]). Also, Non-Redundant Arrays should be, and are often defined as, a particular case of Uniformly Redundant Arrays, but, in common usage and in this thesis, they indicate yet another

Random	
NRA	
Cyclic Difference Sets	• Singer • Hadamard (quadratic residues, twin-primes and *m*-sequences) • Low-density (biquadratic and dilute URAs)

MURAs	
Product Arrays	• PNP • MP • MM • NS
Geometric	
NTHT	
Imperfect masks	

Table 2.1: Summary of array families.

family, with slightly non-ideal properties. In this thesis, the term Pseudo-Noise is not used (but Pseudo-Noise-Product is used) and URA is reserved to twin-prime Hadamard sets so that there be no overlap with *m*-sequences.

General definitions are those of self-supporting masks (one in which all opaque areas are connected at least along a line) and a binary array (an array taking on two values only, typically 0 and 1).

Before entering the details, note that if $\mathbf{A} \times \mathbf{G} = \delta$, also the pair **(1-A, G)** has ideal imaging properties, except for a removable constant sidelobe value. In fact:

$$(\mathbf{1} - \mathbf{A}) \times \mathbf{G} = \mathbf{1} \times \mathbf{G} - \mathbf{A} \times \mathbf{G} = \Sigma \mathbf{G} - N\delta \tag{2.37}$$

The sidelobe value can be removed with an appropriate rescaling of **G**. The negative of a mask, i.e. a mask with open and opaque positions exchanged, is, then, also a mask with ideal imaging properties.

2.4.1 Random array

In the random pinhole array ([10], [11]) holes are placed at random. The decoding array is the array itself, so **G** = **A**. In an array with a total number of positions N_T an arbitrary number $N \leq N_T$ of open positions can be placed. A nice characteristic of the random array is that the density of the holes:

$$\rho = \frac{N}{N_T} \tag{2.38}$$

which is a critical parameter in the determination of the SNR of an aperture (see Chapter 4), can any arbitrary value (unlike some other arrays we will meet later). The peak value of the autocorrelation is *N*,

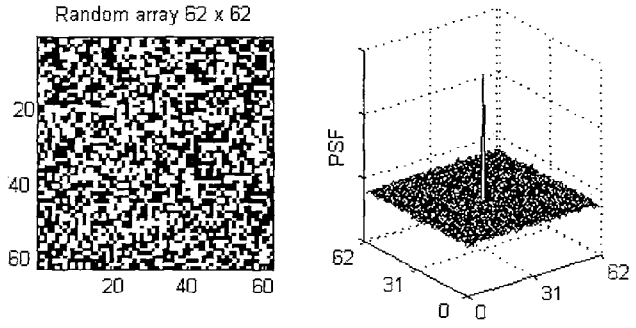


Figure 2.4: 62 × 62 random array and its periodic self-correlation. Note the variation in the sidelobes (inherent noise).

while the off-peak (or sidelobe) value is lower but not constant. In fact, this number is that of the holes still overlapping with other holes after the shift. Since there is no correlation between neighbors, this process is entirely random and the result varies with shift. For this reason, random arrays do not provide ideal imaging properties (see Figure 2.4). This does not necessarily mean that random arrays can not be used without significantly degrading image quality. First, assuming periodic decoding (see section 2.7), the value of any point in the sidelobes is a sample from the same probabilistic distribution: this means that, on the average, sidelobes are flat. Furthermore, since neighbors are not correlated, the sidelobes show no recognizable structures. The effect is that of adding a noisy background, even for an infinite number of counts, but not artifacts. For this reason this "noise" is sometimes called inherent noise, because it is "built in" the mask. The question is whether or not inherent noise is large when compared to statistical noise in typical cases.

Because of independence, shifting the decoding array is equivalent to generating a new one with the exact same number of holes. A given sidelobe value is given by the number n of superpositions with a new random pattern. The number of different patterns of N holes with N_T positions is given by the total number of possible permutations of all positions $N_T!$ divided by the number of permutations of the holes, $N!$, which do not produce a different pattern, and by the number of permutations of the opaque positions, $(N_T\text{-}N)!$, for the same reason. This ratio is the binomial coefficient $\binom{N_T}{N}$. The number of different arrangements leading to the same n is given by the number of possible arrangements of n superpositions over N holes, $\binom{N}{n}$, times the number of different arrangements of the remaining $(N\text{-}n)$ decoding array holes over the $(N_T\text{-}N)$ opaque positions, $\binom{N_T-N}{N-n}$. Of course[5]:

$$\sum_{n=0}^{N}\binom{N}{n}\binom{N_T-N}{N-n}=\binom{N_T}{N} \tag{2.39}$$

and the probability distribution function of n is:

---

[5] Also see Vandermonde's identity $\sum_{m=0}^{n}\binom{r}{m}\binom{s}{n-m}=\binom{r+s}{n}$ for $r+s\geq N$ in ref. [24].

$$p(n) = \frac{\binom{N}{n}\binom{N_T - N}{N - n}}{\binom{N_T}{N}} \tag{2.40}$$

independently of position within the sidelobe. In Appendix B.1 are calculated the average value (ρN) and the standard deviation ($\cong (1-\rho)\sqrt{\rho N}$) of this distribution. Defining the signal as the difference between the peak value and the average sidelobe value and the noise as the standard deviation of the latter, the SNR is:

$$SNR = \frac{N - \rho N}{\sqrt{\rho(1-\rho)^2 N}} = \sqrt{\frac{N}{\rho}} = \sqrt{N_T} \tag{2.41}$$

which increases indefinitely with N_T and is independent of ρ. An infinitely large random array would have perfect imaging properties. In our case, one can see that the 62 × 62 array of Figure 2.4, which has 1860 holes and is thus 48.39% open, has an average sidelobe value of 900, with a standard deviation of about 15.5, as confirmed by a direct evaluation from the autocorrelation function of this particular realization. Since the signal (peak – sidelobe average) is 960, the inherent SNR of this pattern is, as expected, 62. As we shall see in later sections the limit on the SNR imposed by counting statistics is more restrictive, so that this degradation would not be visible.

In conclusion, even if random apertures do not have ideal imaging properties, they are still attractive because they do not contribute structured artifacts to the image and can easily be made of any open fraction and size. Furthermore, even in practical cases they can be large enough so that the inherent SNR is negligible. A major drawback is that very likely they do not have a self-supporting structure.

2.4.2 Non-Redundant Arrays

When the autocorrelation of a binary pattern is taken, the number of hole coincidences at each shift equals the number of holes of the pattern separated by a (vectorial) distance equal to the shift. If this number is 1 for a certain shift, only one hole in the pattern is separated from any other hole by that shift. If this happens for all shifts, the pattern is such that every (vectorial) distance between holes appears only once. For this reason these patterns are called the Non-Redundant Arrays (NRAs). If this were true for all shifts, the self-correlation function would be one everywhere, except for no shift, where a peak of height equal to the number of holes in the pattern is formed. In other words, a delta function is obtained and the

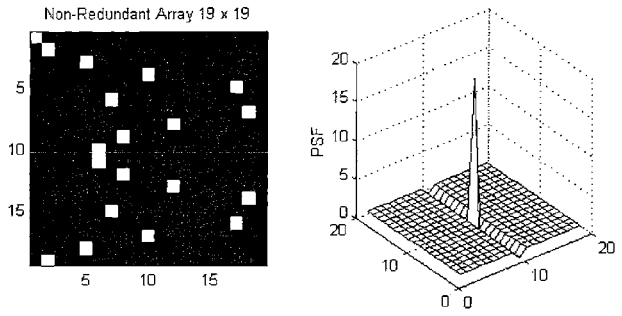


Figure 2.5: 19 × 19 Non-Redundant Array. Note the imperfection in the autocorrelation function.

pattern is an aperture having ideal imaging properties. In practice, sidelobes are only approximately constant, often out to a certain lag, after which they oscillate between 1 and 0 and then eventually drop to 0, when non-periodic correlation is used (see section 2.5). The classic reference on NRAs is ref. [12], which reports a number of patterns with optimal or nearly optimal imaging characteristics. More construction methods are provided in ref. [28], from which the example of Figure 2.5 is taken.

The drawbacks of the NRAs are mainly two: non-optimal imaging properties (which limits the field of view to the region where sidelobes are constant; [3], [6]) and the small number of holes, which results in a small signal throughput ([4], [13]). Nevertheless, they were used in Nuclear Medicine studies ([29]).

2.4.3 Arrays based on cyclic difference sets

The most widely used class of masks relies on cyclic difference sets ([30]). These are a set of N numbers with values between 0 and N_T-1 such that if all possible differences modulo N_T in the set are taken, all solutions between 1 and N_T-1 appear a constant number of times M. In more mathematical terms, a difference set $\mathcal{D}$ is the set of N residues modulo N_T such that for any residue $q \neq 0 \pmod{N_T}$ the congruence $d_i - d_j = q \pmod{N_T}$ with $d_i, d_j \in \mathcal{D}$ has exactly M solutions pairs (d_i, d_j) in $\mathcal{D}$. Cyclic difference sets are characterized and classified by the triplet (N_T, N, M). An example is the set {0, 1, 2, 4} with parameters (7, 4, 2). The sets can also be represented by a binary sequence $\{a_i\}$ of N_T 0s (locations not belonging to the set) and 1s (locations belonging to the set), i.e.:

$$a_i = \begin{cases} 1 & \text{if } i \in \mathcal{D} \\ 0 & \text{if } i \notin \mathcal{D} \end{cases} \tag{2.42}$$

with $i = 0, 1, \ldots, N_T - 1$. In this example the sequence is 1110100. The periodic autocorrelation of these sequences are two-valued:

$$\sum_{i=0}^{N_T-1} a_i a_{\mathrm{mod}(i+l,N_T)} = \begin{cases} N & \text{if } \mathrm{mod}(i+l,N_T) = 0 \\ M = \dfrac{N(N-1)}{N_T - 1} & \text{otherwise} \end{cases} \tag{2.43}$$

This comes as little surprise, because each sidelobe value equals the number of occurrences of a distance, which, by definition of cyclic difference set, is the constant M. The only exception is no shift, in which case the autocorrelation is equal to the total of 1s in the associated sequence, i.e. of elements of the set, N. In this example, one can verify that the peak is 4 and the sidelobes 2.

Two subclasses of cyclic difference sets are Singer sets, for which $N_T = (t^{m+1}-1)/(t-1)$, $N = (t^m-1)/(t-1)$ and $M = (t^{m-1}-1)/(t-1)$, where t is a prime number, and Hadamard sets, for which $N_T = 4t-1$, $N = 2t-1$ and $M = t-1$, where t is an integer. Hadamard sets are at the basis of Pseudo-Noise (PN) sequences and can be classified in ([3]):

Quadratic residues: the set is given by the squares, mod(N_T), of the first $(N_T+1)/2$ integers, with N_T prime.

Twin primes: a set is twin prime if N_T is the product of two prime numbers whose difference is 2.

m-*sequences*: sets for which $N_T = 2^m-1$ with m integer greater than 1. These also fall under the definition of Singer sets for $t = 2$.

Some sets may belong to more than a subclass. For example, our sample sequence is both a quadratic residue and an m-sequence. In fact, the squares of the first $(7+1)/2 = 4$ integers are 0, 1, 4, 9 and 16, which, mod(7), are 0, 1, 4, 2 and 2. Furthermore, if $m = 4$, $2^4-1 = 7 = N_T$.

From the definition, Singer sets can be seen to have density $N / N_T \cong 1 / t$. Hadamard sets all have a density of about 50%. Hadamard sets are related to Hadamard matrices, orthogonal matrices ($\mathbf{H}^{-1} = \mathbf{H}^{\mathrm{T}}$, i.e. row (column) vectors are mutually orthogonal) whose elements are either 1 or –1. Hadamard matrices have a close relationship with m-sequences. An m-sequence can be placed as the first row of a cyclic matrix, by definition one in which a line is obtained from the previous by shifting all elements one place to the right and moving the last element at the right to the left. If a column of ones and a row of zeros is added to the matrix so obtained and the resulting matrix is rescaled between one and minus one, a

Hadamard matrix is obtained ([26]). The equivalence of URAs and Hadamard coded apertures is also treated in ref. [27] and [14].

Cyclic difference sets generate 1d sequences that, for our ends, need to be folded in two dimensional arrays. For rectangular arrays this is possible only if N_T is not prime, which excludes quadratic residue and part of the *m*-sequences. A first way of folding a 1d sequence in a 2d sequence is to place the elements along a diagonal continuing at the opposite side when an edge is reached (Figure 2.6a), [31]). While with this method the periodic self-correlation of the 2d array is immediately a δ function, it works only if N_T can be factorized in the product of two mutually prime numbers. A more straightforward way is to arrange the sequence by rows and columns, for instance lexicographically. However, one can not decode by simply taking **G** = **A** and using periodic correlation (see section 2.7). This is because when the self-correlation is calculated, while all elements moving, say, one place to the right are displaced by one position on the decoding array, the last column must be reinserted to the left, which is equivalent to a much longer displacement, equal to the size of the mask. Since not all elements are displaced by the same amount, correlation properties are lost. The decoding array needs to be defined in the way seen in Figure 2.6b, but this requires that the decoding array be larger than the mask. An advantage of this folding is that the dimensions need not be mutually prime.

URA: Uniformly Redundant Arrays

The Uniformly Redundant Arrays (URAs) are the sequences indicated as pseudo-noise by Calabro and Wolf in 1968 ([25]). Their application to imaging was expounded in 1978 by Fenimore and Cannon ([13]). The name comes from the property that all separations between holes in the pattern occur a constant number of times (*M*). If *M* were 1, we would be reduced to the case of the NRAs. In all other cases all distances appear a constant number of times, hence the uniform redundancy. Of course, this

1	16	31	11	26	6	21
22	2	17	32	12	27	7
8	23	3	18	33	13	28
29	9	24	4	19	34	14
15	30	10	25	5	20	35

a)

0	1	2	3	4	5	6	7	8
5	6	7	8	9	10	11	12	13
10	11	12	13	14	0	1	2	3
0	1	2	3	4	5	6	7	8
5	6	7	8	9	10	11	12	13

b)

Figure 2.6: a) diagonal folding of a 35-element 1d array in a 5 × 7 2d array. b) Decoding array associated to a 3 × 5 array folded by rows. Note how the last column and row were left out to avoid ambiguities in the decoding (see aliasing in section 2.6).

property is common to all arrays coming from cyclic different sets, but in common usage the term URA came to indicate twin-prime Hadamard arrays. Fenimore and Cannon give a procedure to generate directly a 2d array bypassing folding:

$$A_{i,j} = \begin{cases} 0 & \text{if } i = 0 \\ 1 & \text{if } j = 0,\ i \neq 0 \\ 1 & \text{if } c_r(i)\, c_s(j) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.44}$$

where

$$c_r(i) = \begin{cases} 1 & \text{if there exists an integer } x,\ 1 \leq x < r \\ & \text{such that } i = \mathrm{mod}_r\, x^2 \\ -1 & \text{otherwise} \end{cases} \tag{2.45}$$

and $| r - s | = 2$, with r and s prime. The positive coefficients $\mathbf{c}_r$ are called quadratic residues modulo r. $\mathbf{c}_s$ is defined similarly to $\mathbf{c}_r$. The array is generated as an $r \times s$ 2d array. By definition, it is a twin-prime array. In fact, it can also be generated ordering along the diagonal a twin-prime $r \times s$ sequence.

For a URA:

$$M = \frac{N}{2} \quad \text{and} \quad N = \frac{N_T + 1}{2} \tag{2.46}$$

so that the density is about 50%, as for any Hadamard set.

From the definition of cyclic difference set, the decoding array is the mask array itself, except for a linear rescaling (see section2.7). It is important to stress once again that the autocorrelation properties

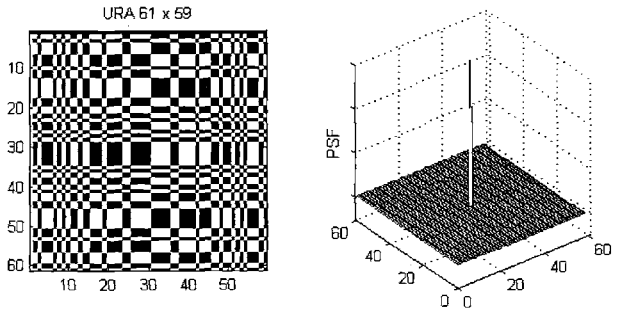


Figure 2.7: 61 × 59 Uniformly Redundant Array. Note the perfect autocorrelation function.

m	Length (2^m-1)	Size	m	Length (2^m-1)	Size	m	Length (2^m-1)	Size
1	1	1	6	63	9×7	11	2047	89×23
2	3	3×1	7	127	127×1	12	4095	65×63
3	7	7×1	8	255	17×15	13	8191	8191×1
4	15	5×3	9	511	73×7	14	16383	129×127
5	31	31×1	10	1023	33×31	15	32767	217×151

Table 2.2: sizes of *m*-sequence arrays available by folding along the diagonal. The size closest to a square is listed. Note that sequences with $m = 2, 3, 5, 7$ and 13 can not be folded in a 2d array because their length is a prime number.

are ideal only for periodic correlation. The reason is clear from the construction procedure and ultimately descends from the definition of distance in a cyclic difference, which is a distance modulo N_T.

Finally, it is worth mentioning that a subset of the quadratic residue arrays (those in the form $N_T = 12s+7 = 4(3t+2)-1$ with N_T prime) can be folded in hexagonal arrays ([3], [14]).

m-sequences

An *m*-sequence (maximal-length linear shift register sequence) is built from an irreducible (or primitive) polynomial of degree m of coefficients $p_0, p_1, \ldots p_{m-1}$, where p_i is either 0 or 1, $\forall\ i$ ([32]). The i-th element of the sequence is given by:

$$a_{i+m} = \sum_{j=0}^{m-1} p_j a_{i+j} \quad i = 0, 1, ..., 2^m - m - 2 \tag{2.47}$$

where the summation is a summation modulo 2 (so $1 + 1 = 0$) and $a_0, a_1, \ldots, a_{m-1}$ are chosen arbitrarily (but are still either 0 or 1), each choice resulting simply in a different cyclic shift of the final sequence. Sequences have lengths restricted to 2^m-1 and a number of open positions 2^{m-1}-1, for a density of about 50%, as expected for a Hadamard sequence. An alternative construction method generating directly a 2d array is given in ref. [25], where *m*-sequences are also called Gordon arrays. If m is even, the length of the sequence can be factorized in the product $(2^{m/2}+1) \times (2^{m/2}-1)$, which is an almost square array.

An example is given in Figure 2.8, where a 65 × 63 *m*-sequence is shown with its self-correlation function. Its structure is very different from that of a twin prime array, a property of potential interest in both the study of near-field artifacts and 3d laminography ([33]). Unfortunately, for this class of masks

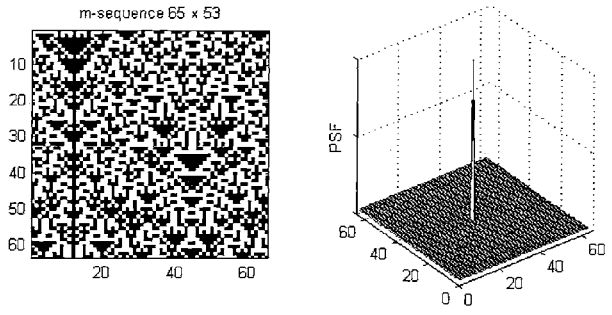


Figure 2.8: 65 × 63 m-sequence. m = 12. Note the perfect autocorrelation function.

not many pattern sizes exist because the length is constrained to 2^m-1. In Table 2.2 are shown the possible patterns with $m \leq 15$. It is also possible to fold m-sequences in such a way that a self-supporting array is obtained ([34]).

Low density arrays: biquadratic and dilute URAs

Some low density arrays still have perfect autocorrelation properties and yet are not part of the above categories. They are the biquadratic arrays and the dilute URAs.

Biquadratic arrays have a 25% open fraction ([35]). Given an odd integer t such that $N_T = 4t^2+1$ is prime, a 1d biquadratic sequence a_i is generated with the following rule:

$$a_i = \begin{array}{ll} 0 & \text{if } i = 0 \\ 1 & \text{if } i = \text{mod}(j^4, N_T) \text{ for any } 0 < j < n \text{ and } i \neq 0 \end{array} \qquad (2.48)$$

For these arrays $N = (N_T\text{-}1)/4$. Unfortunately the restrictions on N_T allow only a few possible lengths: 5, 37, 101, 197, 677, 2917, 4357, 5477, ...

The dilute URAs were found with an extensive computer search ([36]). Their lengths are given by $N_T = N(N\text{-}1) + 1$. The search found 1 code of length 21, 5 of length 31, at least 6 of length 57 and at least one of length 73. No codes of length 43 were found. For N = 2, 3 and 4 codes previously known as Barker codes were found. Since no generation rule is available, arrays are simply listed.

Except for the 57 dilute URA, the length of all these sequences is prime. Nevertheless, these sequences are useful in the generation of 2d arrays if they are taken as starting 1d sequences for product arrays (see §2.4.5).

2.4.4 MURA: Modified Uniformly Redundant Array

Quadratic Residue Arrays are an extension of the one-dimensional quadratic residue sequences of §2.4.3 ([25]). These arrays, of area $p \times q$, with p and q prime, have different properties depending on whether $q - p = 0$, 2, 4, or 6. For twin-prime arrays (URAs) $|q - p| = 2$. This is the only case in which the autocorrelation of the array is a δ function. However, also arrays for which $q - p = 0$ have interesting properties, the first of which being that they are square. The arrays are generated with the same algorithm used for the URAs, so:

$$A_{i,j} = \begin{cases} 0 & \text{if } i = 0 \\ 1 & \text{if } j = 0,\ i \neq 0 \\ 1 & \text{if } c_p(i)\, c_p(j) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.49}$$

where

$$c_p(i) = \begin{cases} 1 & \text{if there exists an integer } x,\ 1 \leq x < p \\ & \text{such that } i = \mathrm{mod}_p\, x^2 \\ -1 & \text{otherwise} \end{cases} \tag{2.50}$$

However, **G** can not be equal to **A** because the autocorrelation is not a δ function. Gottesman and Fenimore ([14]) pointed out that a slight modification in the definition of **A** gives the **G** resulting in ideal imaging properties ($\mathbf{G} \otimes \mathbf{A} = \delta$). The decoding function is now[6]:

$$G_{i,j} = \begin{cases} 1 & \text{if } i \oplus j = 0 \\ 1 & \text{if } A_{i,j} = 1,\ i \oplus j \neq 0 \\ 0 & \text{if } A_{i,j} = 0,\ i \oplus j \neq 0 \end{cases} \tag{2.51}$$

The only difference is in the element $i \oplus j = 0$ which is now 1 rather than 0. The only limitation on the parameters is now that p be prime. So square arrays of side 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89 and 97 are possible, when masks with less than 10000 elements are considered.

As it can be seen in Figure 2.9 these patterns are invariant for a 180° rotation about the center. For a 90° rotation some arrays, which we will call symmetric, are also invariant, while others, which we will

[6] In the original paper Gottesman and Fenimore propose balanced decoding and indicate a coefficient –1 in place of 0. See the discussion of Chapter 4 for details.

call anti-symmetric, exchange open and closed positions. All arrays however, have the same open fraction. Since:

$$M = \frac{N}{2} \quad \text{and} \quad N = \frac{N_T - 1}{2} \tag{2.52}$$

the open fraction is $N / N_T \cong 50\%$ as for URAs.

MURAs can also be generated as 1d sequences. Provided $N_T = 4t+1$ ($t \in \mathbb{N}$) and prime, the binary sequence:

$$a_i = \begin{cases} 0 & \text{if } i = 0 \\ 1 & \text{if } i = \text{mod}(j^2, N_T) \text{ for any } 0 \le j < n \text{ and } i \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.53}$$

and the decoding array:

$$g_i = \begin{cases} 1 & \text{if } i = 0 \\ 1 & \text{if } a_i = 1,\ i \neq 0 \\ 0 & \text{if } a_i = 0,\ i \neq 0 \end{cases} \tag{2.54}$$

are such that $\mathbf{a} \otimes \mathbf{g} = \delta$. The only difference between **a** and **g** is the first element of the sequence.

Like URAs, MURAs can be folded in hexagonal arrays.

2.4.5 *Product Arrays*

Another class of arrays with ideal correlation properties derives from the implementation of the following ([25]):

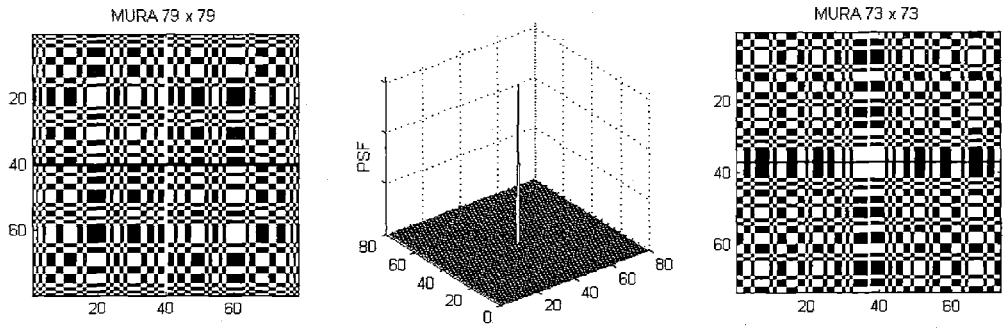


Figure 2.9: MURA patterns. While the 79×79 is anti-symmetric about its center, the 73×73 is symmetric. To show symmetry, the pattern was periodically shifted so that lines $i = 0$ and $j = 0$ are at the center.

THEOREM: If **A** is a 2d array formed by taking the term-to-term product of two 1d sequences **b** and **c** ($\mathbf{A} = \mathbf{b}\,\mathbf{c}^T$), then the 2d auto-correlation of **A** is the product of the individual autocorrelations of the sequences, i.e.:

$$\mathbf{A} \otimes \mathbf{A} = (\mathbf{b} \otimes \mathbf{b})\,(\mathbf{c} \otimes \mathbf{c})^T \tag{2.55}$$

where vectors are column vectors and ^{T} indicates transposition.□

If **b** and **c** are perfect arrays, **b** ⊗ **b** and **c** ⊗ **c** are 1d δ functions, their product is a 2d δ function, so **A** is also a perfect array, which is named, after its construction procedure, a product array. Moreover, the theorem can be extended to cross-correlations as well ([25]):

THEOREM: If **A** and **B** are 2d arrays formed by taking the term-to-term product of two 1d sequences ($\mathbf{A} = \mathbf{c}\,\mathbf{d}^T$ and $\mathbf{B} = \mathbf{e}\,\mathbf{f}^T$), where **c** and **e**, and **d** and **f**, have the same length, then the 2d cross-correlation of **A** and **B** is the product of the individual cross-correlations of the 1d sequences, i.e.:

$$\mathbf{A} \otimes \mathbf{B} = (\mathbf{c} \otimes \mathbf{e})\,(\mathbf{d} \otimes \mathbf{f})^T.\square \tag{2.56}$$

This theorem reduces the 2d problem to two separate 1d problems. 2d arrays can be generated from any pair of 1d sequences having perfect cross correlation, i.e. such that **c** ⊗ **e** = δ and **d** ⊗ **f** = δ, where δ is a 1d δ function. Note that non-zero sidelobe values of the 1d cross-correlation function can not be ignored because such values would, upon external multiplication of the 1d arrays, replicate the spike present in the other array in all rows (columns) and the 2d cross-correlation function would then look like a cross with a peak at the center. 1d decoding arrays must be scaled so that the 1d cross-correlation is truly a δ. This precludes decoding any array (except for a pinhole) with itself: rescaled versions of the decoding array must be used in place of sequences of 0s and 1s. If the 1d sequences are URAs or MURAs, the associated decoding sequences are made of 1s and –1s. In fact, the value of the sidelobes is given by the number of holes that still superimpose after shift (in this case $N/2$) plus the number of holes that do not (of course also $N/2$) times the value associated to closed positions in the decoding array. Setting this value to –1 is equivalent to setting the sidelobe value to 0, as is necessary for the 1d sequences of product arrays.

No relationship is constraining the size of the 2d array other than those limiting the originating 1d sequences. 2d arrays take different names depending on the 1d sequences chosen.

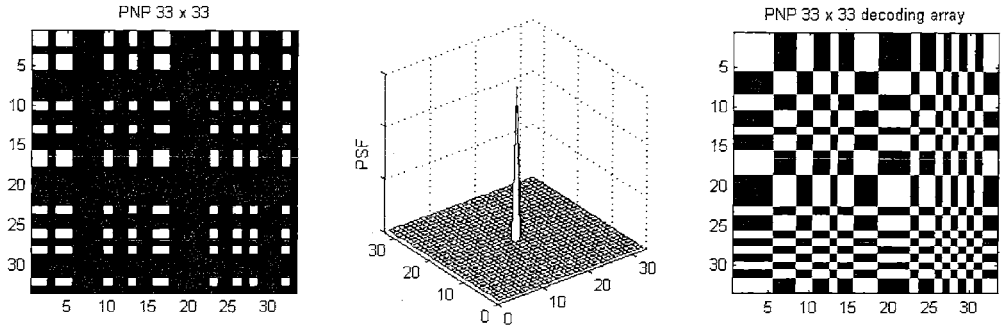


Figure 2.10: 33×33 product array, its cross correlation function, and the associated decoding array. For the decoding array, white represents +1 and black −1.

Pseudo-Noise Product arrays: PNP[7].

If the 1d sequences are quadratic residue sequences, the product array takes the name of PNP or Pseudo-Noise Product ([37]). The correct decoding array is obtained from the point-by-point multiplication of the properly scaled 1d decoding sequences coefficients. The 2d array can be rescaled after its calculation in any arbitrary way without perturbing autocorrelation properties, the only effect being that of adding a pedestal value. Peak and sidelobe values can so be modified arbitrarily.

A nice characteristic of these arrays is that they are self-supporting. Their open fraction, being the product of two half-open sequences, is about 25%. In Figure 2.10 is the example of a 33×33 array.

M-P and M-M arrays

If a 1d MURA sequence is used in place of a PN sequence, an array of the M-P family is obtained ([38]). If both 1d sequences are 1d MURAs, an M-M array is obtained. The properties of these families are completely similar to those of the PNP arrays. However, they are important in that they extend the number of available pattern sizes.

NS arrays

A second cross-correlation theorem can be used to generate low open-fraction 1d arrays ([39]).

[7] The authors of ref. [35] seem to indicate with PN a quadratic residue sequence.

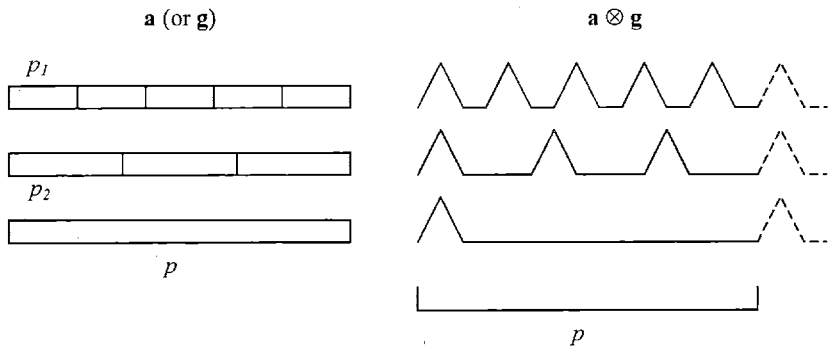


Figure 2.11: formation of a NS array and its cross-correlation function.

THEOREM: If $\mathbf{a}_x$ are $x = 1, 2, \ldots, s$ different 1d sequences, each of order p_x, mutually prime, and $\mathbf{g}_x$ the corresponding decoding functions, then the new sequence, of order $p = \prod_{x=1}^{s} p_x$ defined by:

$$a_p = \prod_{x=1}^{s} a_x(\mathrm{mod}_{p_x}\, i),\ g_p = \prod_{x=1}^{s} g_x(\mathrm{mod}_{p_x}\, i) \tag{2.57}$$

has the cross-correlation function:

$$R_p = \prod_{x=1}^{s} R_x \tag{2.58}$$

where R_x is the cross-correlation function of the pair $(\mathbf{a}_x, \mathbf{g}_x)$.□

It is still important that $\mathbf{g}_x$ be scaled so that the sidelobes of the cross-correlation be 0. In practice, a low open fraction 1d array is generated by taking two (or more, in general s) 1d arrays, with relatively prime lengths, p_1 and p_2. The first array (and decoding array) is replicated p_2 times while the second is replicated p_1 times, then a point-by-point multiplication is made. The new sequence, which is long $p = p_1 \times p_2$, has a cross correlation function which is the point-by-point product of the individual cross-corrrelations. These are δ functions within the period p_1 and p_2, respectively. When the periods are replicated, so are the δ functions in the individual cross-correlations. From application of the theorem above, the cross-correlation of the product array is the product of the cross-correlations. Since p_1 and p_2 are prime to each other, only a δ for every period p survives multiplication and the product array has

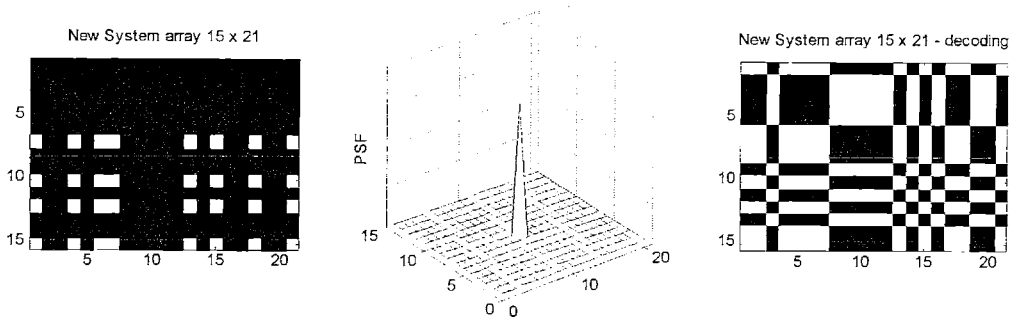


Figure 2.12: 15 × 21 NS array generated from two NS 1d sequences. These are formed from a 3 and a 7 URA 1d sequence and a 3 URA and a 5 MURA 1d sequence. The density of this pattern is 10.2%.

perfect correlation properties over a period p (see Figure 2.11), just like the initial sequences had perfect properties over p_1 and p_2. In symbols:

$$R_p(k) = \prod_{x=1}^{s} R_x(k) = \begin{cases} n_x & \mathrm{mod}_p(k) = 0 \\ 0 & \mathrm{mod}_p(k) \neq 0 \end{cases} \tag{2.59}$$

with n_x the number of holes in each of the original sequences. The open fraction of each sequence is $T_x = n_x / p_x$ and the open fraction of the new pattern is

$$T_p = \prod_{x=1}^{s} T_x = \prod_{x=1}^{s} \frac{n_x}{p_x} = \frac{n_p}{p} \tag{2.60}$$

The new 1d sequences can then be combined with the methods of the previous sections to give a 2d array of open fraction equal to the product of the open fraction of the two 1d sequences used. These 2d arrays were not given a specific name but are referred to as the "new system" by their discoverer ([39]). We will use the initials NS. They are a further generalization of the product arrays of previous paragraphs. They extend the number of sizes available, are all self-supporting, and, more importantly, the values of the open fraction are not constrained around 25% but range from a few percents to 30%. Arrays of the same size can be constructed in different ways, each leading to a different open fraction.

2.4.6 Geometric arrays

Geometric apertures are recognizable from regularities in their pattern. Several shapes have been suggested ([40]-[42]), but final designs seem to focus on the L and the X type ([42]; these are design I and

III of ref. [40]). Two examples are shown in Figure 2.13. The autocorrelation function is not ideal, but it is possible to define three-level decoding arrays so that the cross-correlation is ideal. For the L family the three levels are:

$$\begin{array}{rl} \alpha = [5(n\text{-}1)\text{-}n^2] \,/\, (n\text{-}2) & \text{at the top left} \\ 1 & \text{on the first row and column (except top left)} \\ \beta = -1 \,/\, (n\text{-}2) & \text{elsewhere} \end{array} \qquad (2.61)$$

where n is the side of the array, which is square. The array density is $2(n\text{-}1) / n^2 \sim 2 / n$. For the X family they are:

$$\begin{array}{rl} \alpha = [-11+9n\text{-}n^2] \,/\, (n\text{-}4) & \text{at the top left} \\ 1 & \text{on the first row and column and diagonals (except top left)} \\ \beta = -3 \,/\, (n\text{-}4) & \text{elsewhere} \end{array} \qquad (2.62)$$

The density is exactly twice that of the L family: $4(n\text{-}1) / n^2 \sim 4 / n$. For L arrays there is no restriction on n, while X arrays need n odd. Other densities are achieved by adding rows and lines or eliminating a diagonal: these are the geometric families not mentioned here.

Unfortunately these apertures do not provide the same SNR of other patterns, noticeably the URAs. Furthermore, as we shall see in Chapter 4, noise does not affect the image uniformly but some points show a SNR much lower than others, so that image quality depends on the location of sources on the field of view. This is because the decoding coefficients are not unimodular (i.e. do not have all the same magnitude) as is for product arrays and cyclic difference set sequences.

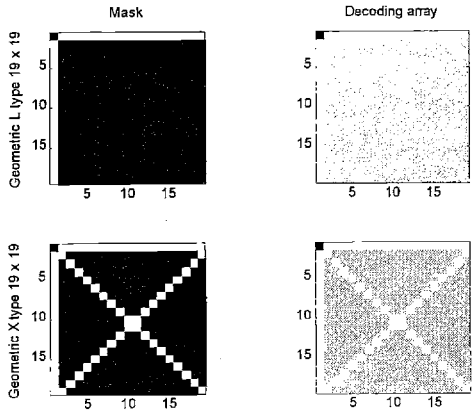


Figure 2.13: geometric coded apertures. Top left: 19×19 L array. Top right: associated decoding array. Bottom left: 19×19 X array. Bottom right: associated decoding array. The decoding functions have three levels (see text for exact values) indicated in white, black and gray.

2.4.7 *NTHT arrays*

So far arrays have been treated as dimensionless sequences of numbers. **A** was assumed to be a square grid of open and closed positions. This is not necessary at all and actually leads to an imprecision most easily seen switching from a discrete representation to the underlying continuous-space representation. In the following the finite size of the hole is shown to prevent the correlation of **A** and **G** to be a δ, which is true only in the sense of a discrete array representation. Building on this idea a new family of arrays is discovered.

Let A be a continuous-space 2d function representing a mask with perfect autocorrelation properties, as shown in any of the previous figures, based on the 2d array $\mathbf{A}_{i,j}$. A can be written as the convolution:

$$A = A_\delta * S_m \tag{2.63}$$

where S_m is a 2d function describing the shape of a mask hole (assumed all equal) and A_δ is a sum of displaced δ functions (continuous variable) placed at the centers of the holes:

$$A_\delta(\vec{r}) = \sum_{i,j} \mathbf{A}_{i,j}\delta(\vec{r} - \vec{r}_{i,j}) \tag{2.64}$$

The vectors $\vec{r}_{i,j}$ form a rectangular array of positions at some of which is present a hole, according to **A**. S_m is non-zero only on a small area around $\vec{r}_{i,j}$. In the examples above it is a square of side equal to the spacing of the $\vec{r}_{i,j}$ s. In the following imaging properties are shown not to be disrupted if the shape of the holes is changed (for a frequency-space derivation of the same result see [43]): S_m turns out to contribute only a constant. This observation is at the basis of an extension of any of the masks presented above, the No-Two-Holes-Touching masks ([44]).

The continuous function describing the photon distribution projected on the detector is, in the far-field approximation:

$$R(\vec{r}) = A(\vec{r}) = \iint A'_\delta(\vec{\xi})S'_m(\vec{r} - \vec{\xi})d^2\vec{\xi} \tag{2.65}$$

where a point source has been assumed without compromising generality because of linearity and A'_δ and S'_m are defined, respectively, as the projection of the mask holes and their shape on the detector:

$$A'_\delta(\vec{r}) = A_\delta(\frac{a}{z}\vec{r}), \qquad S'_m(\vec{r}) = S_m(\frac{a}{z}\vec{r}) \tag{2.66}$$

The detector samples this projection on a rectangular lattice $\vec{r}_{r,s}$ by integrating on an area surrounding each node of the lattice described by the detector pixel shape S_p, also assumed all equal. The number of counts are stored in the matrix $\mathbf{R}_{r,s}$, the digitized projection according to:

$$\mathbf{R}_{r,s} = \iint_{\vec{r}} R(\vec{r}) S_p(\vec{r} - \vec{r}_{r,s}) d^2\vec{r} = \iint_{\vec{r}} S_p(\vec{r} - \vec{r}_{r,s}) \iint_{\vec{\xi}} A'_\delta(\vec{\xi}) S'_m(\vec{r} - \vec{\xi}) d^2\vec{\xi} d^2\vec{r} \tag{2.67}$$

and expressing A_δ in terms of $\mathbf{A}$:

$$\mathbf{R}_{r,s} = \sum_{i,j} \mathbf{A}_{i,j} \iint_{\vec{r}} S_p(\vec{r} - \vec{r}_{r,s}) S'_m(\vec{r} - \vec{r}_{i,j}) d^2\vec{r} = \sum_{i,j} \mathbf{A}_{i,j} \mathbf{F}_{r-i,s-j} \tag{2.68}$$

where:

$$\mathbf{F}_{r-i,s-j} = \iint_{\vec{r}} S_p(\vec{r} - \vec{r}_{r,s}) S'_m(\vec{r} - \vec{r}_{i,j})\, d^2\vec{r} = S_p \times S'_m \tag{2.69}$$

is the form factor and provides the bridge between the analog signal and its digital representation. In the hypotheses that the two grids $\vec{r}_{i,j}$ and $\vec{r}_{r,s}$ have the same pitch, $\mathbf{R}_{r,s}$ is the discrete convolution:

$$\mathbf{R} = \mathbf{A} * \mathbf{F} \tag{2.70}$$

If the projection of a mask hole covers at most a detector pixel:

$$\mathbf{F}_{r-i,s-j} = f\, \delta(r - i, s - j) \tag{2.71}$$

where f is the area of the superposition of hole projection and detector pixel. So the form factor is zero everywhere except at one location. Substitution in eq. (2.68) gives:

$$\mathbf{R}_{r,s} = \sum_{i,j} \mathbf{A}_{i,j} f\, \delta(r - i, s - j) = f\, \mathbf{A}_{r,s} \tag{2.72}$$

which can be decoded by discrete correlation with the decoding array $\mathbf{G}$:

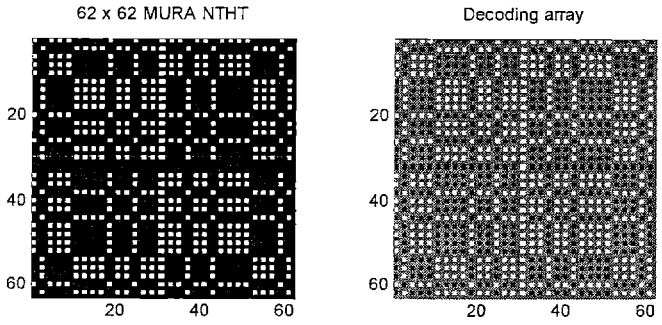


Figure 2.14: NTHT array derived from a 31 × 31 MURA. An opaque row and column was inserted between all rows an d columns. The Decoding array is three valued: 1 white, 0 gray, -1 black.

$$\hat{\mathbf{O}}_{u,v} = \mathbf{R} \otimes \mathbf{G} = \sum_{r,s} f\mathbf{A}_{r,s}\mathbf{G}_{r\oplus u, s\oplus v} = f\,\delta(u,v) \tag{2.73}$$

which is the point source. The only effect of the hole shape is to multiply the image by a constant. This ultimately comes from the hypothesis made that the projection of one hole fall on one pixel only[8]. If in any of the arrays so far encountered the holes are reduced by adding a border around them, we still have a valid array. All holes become isolated, hence the name No-Two-Holes-Touching (NTHT) and the array also becomes self-supporting. If the added border is such that the side of the hole is reduced by a factor $1/e$, the operation is equivalent to adding e-1 opaque rows and columns of size $1/e$ between all mask positions. The open fraction is so reduced by a factor e^2.

From the proof above, the array has still ideal properties if sampled according to its original size, i.e. when the added rows and columns are not seen as new mask pixels, but just as shape factor. However, the mask can be thought of as a completely new array each pixel of which is sampled in the projection, including the opaque new ones. The decoding array has to be modified to match the new dimensions of the recorded data. This is done by adding lines of 0s corresponding to the new rows and columns. If the original decoding array was scaled so that sidelobes in the PSF are zero, the NTHT array has indeed ideal properties. In fact, for the shifts of **G** that superimpose pixels of the original mask and decoding arrays, the cross-correlation is always 0 except for no shift, which gives the peak value N. For all other shifts all mask holes fall on the newly added zeros of the decoding function, giving again a sidelobe value of 0 and preserving ideality. As for geometric arrays, now the decoding array is a three-leveled function.

[8] This hypothesis will be relaxed in section 2.7 where a more general case is discussed.

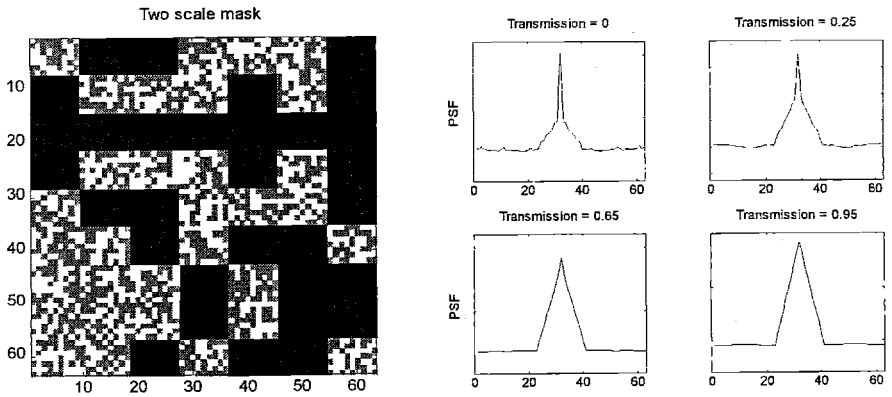
2.4.8 Imperfect masks: two spatial scales

A very interesting design was proposed by Skinner and Grindlay ([45]). In X-ray astronomy detectors and masks have to cover as wide an energy range as possible. Masks designed for low-energy imaging typically have fine details to achieve good resolution. They are lost at higher energy, because photons are more penetrating and, in addition, can not be located as precisely at the detector. The idea is to use a mask which is the superposition of two masks with two spatial scales, a finer one for low energy and a coarser one for high energy. The low energy mask is thin enough to become almost transparent at high energies (Figure 2.15a). If the coarser mask has perfect imaging properties, high-energy imaging has poor resolution, because the mask is coarse, but still perfect properties. At low energy resolution is higher, because the finer mask is opaque, but the PSF is not perfect because the coarser mask superimposed cancels some of its parts.

The PSF as a function of energy is in Figure 2.15b. For low energy, a broad peak due to the coarse mask is surmounted by a thinner peak, due to the finer mask. Since in this case the finer mask is a random array, sidelobes are not flat. As transmission through the fine mask increases with energy the broader peak increases until it is the only structure left, indicating inferior resolution (see section 2.6).

2.5 Coded aperture camera geometries

From the discussion of section 2.4 it is clear that perfect imaging is possible for masks of finite



a) b)

Figure 2.15: two scale mask. a) A 9×7 *m*-sequence is superimposed to a random array, partially opaque. b) Section of the PSF of this mask for increasing transmission.

size only for arrays that are decoded with a periodic correlation. It is also clear that perfect imaging is possible only if complete mask patterns are projected on the detector. This section presents the geometrical arrangements of mask and detector that realize these prerequisites ([13]). In the following we assume a point source located at infinity, so that the projection of the mask on the detector has the same size as the mask itself. The argument is easy to extend to near-field sources.

The most straightforward arrangement is that of Figure 2.16a. The mask and detector have the same size. This configuration is called the box camera ([4], [46]). In this case only a source on the instrument axis can cast a complete mask pattern on the detector. The Fully-Coded Field of View (FCFV or simply FoV) is a point. All other sources still project some part of the mask on the detector (they are part of the Partially-Coded Field of View (PCFV)) and can not be reconstructed perfectly. Of course, one could make a smaller mask, but other solutions turn out to be more advantageous. A larger detector is certainly a possible solution (Figure 2.16b), but very often fabrication issues or cost set a limit on detector size. In these cases it is more convenient to instead enlarge the mask, by replicating it in a 2×2 arrangement called mosaic (Figure 2.16c). All sources within the FCFV still project an entire mask pattern on the detector, but the pattern shifts are different, depending on source position. The FCFV is enlarged at the expense of a larger mask. Note that points at the very border of the FCFV all project the same pattern. For example, the point that projects the top left corner of the mosaicked mask onto the top left corner of the detector projects the same pattern shift as the point projecting the center of the mosaic on the same detector corner. The two sources, despite having different locations would be reconstructed at the same point. This problem, called aliasing, is why replication of the mask can not go beyond 2×2, indeed not even reach 2×2 but should, rather, leave out a row and a column. This said, in the calculations

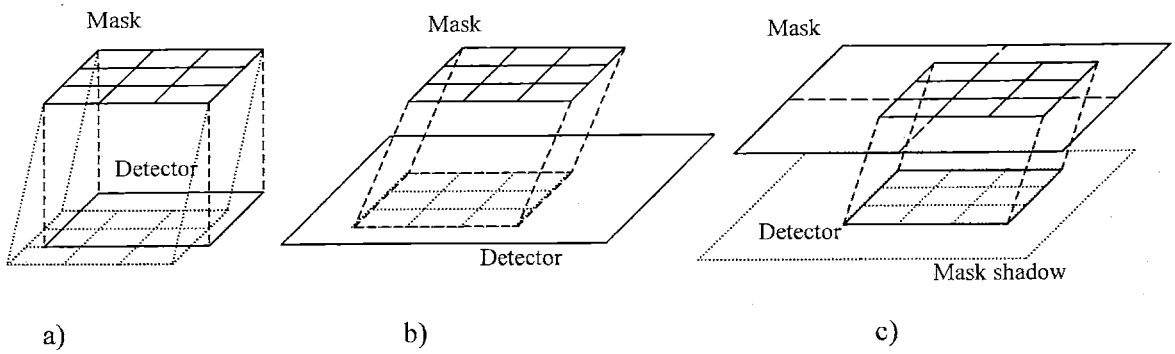


Figure 2.16: three possible coded aperture camera geometries. a) Mask and detector have the same size. b) The detector is larger than the mask. c) The mask is larger than the detector. It is a 2×2 mosaick of the basic pattern used in a) and b). Note that parts of the mask projection miss the detector. This does not matter as long as the detector captures some full period of the mask. Point source at infinity assumed in all cases.

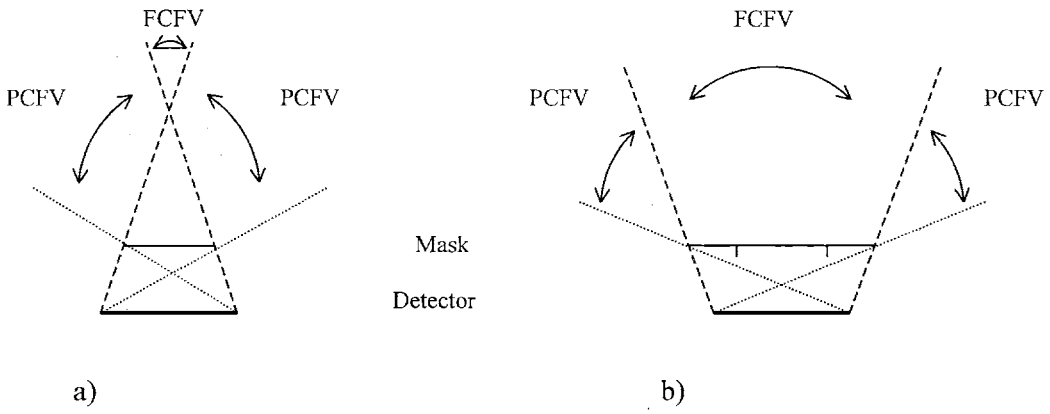


Figure 2.17: Fully-coded field of view (FCFV) and Partially-Coded Field of View (PCFV) for a) non-cyclic and b) mosaicked geometry. Source assumed at infinity. Detector assumed larger than a mask period.

that follow we will always assume for simplicity that the mask is a complete 2×2 array.

The problem with the sources in the PCFV is not eliminated with the second two arrangements. The simplest solution is to shield these sources or introduce collimators that exclude them from seeing the detector, but this is not always possible, a notable example being astronomy. In these applications, the effect of partial encoding combines dangerously with aliasing to a degree that the partial encoding of the box camera is a lesser evil ([47]), even if one has to resort to Wiener filtering or maximum entropy algorithms to reconstruct the image ([48]). This is not surprising, because faulty encoding due to discontinuities in the detector can also be overcome ([49]). In planar Nuclear Medicine studies the PCFV is very easy to shield, and therefore we do not need to consider the box camera any longer. A geometric representation of the extension of FCFV and PCFV is in Figure 2.17.

2.6 Field of View and resolution

In the previous section, the geometric configuration of the coded aperture was shown to determine the field of view of the camera. A second geometric parameter of great importance is the geometric resolution, i.e. the resolution of the system due to its geometric design, assuming the detector to be ideal. Actual images have worse resolution, the system resolution, which also includes the effects of a non-ideal detector, such as intrinsic point spread function and sampling. This extension is fairly complex and is postponed to Chapter 7.

Field of view and geometric resolution are strictly related to each other and to mask and detector parameters. The analysis can be carried out in 1d because generalization to 2d is immediate. Two near-

field configurations are examined. The first resembles that of the box camera: a single period of the mask is projected on the detector. In this case the whole projection must be cast on the detector. The mask, of side d_m, is smaller than the detector, of side d_d. It is convenient to define the magnification coefficient as the ratio of the projection of the mask to the size of the mask itself:

$$m = \frac{d_m \dfrac{a+b}{a}}{d_m} = 1 + \frac{b}{a} \tag{2.74}$$

where, consistently with previous usage, a is the object-to-mask distance and b the mask-to-detector distance. m is always greater than 1, a value which is approached as the object is moved away from the camera. Note that this is not the magnification of the pinhole camera m_p, which was defined as the ratio of size of the projection of the object to the size of the object itself. From eq. (2.5), the relation between the two coefficients is:

$$m = m_p + 1 \tag{2.75}$$

With this definition the mask projection has size $m\,d_m$, which leaves a space $(d_d - m\,d_m)/2$ available on both sides of the center to shift the mask. By looking at the shaded triangles of Figure 2.18a:

$$\frac{H}{2} = \frac{\dfrac{d_d - m d_m}{2}}{b} \cdot a \tag{2.76}$$

so that the field of view is:

$$H = \frac{d_d - m d_m}{m-1} = \frac{d_d}{m-1} - \frac{m d_m}{m-1} \tag{2.77}$$

In a box camera $m\,d_m = d_d$ and $FoV = 0$ as already discussed.

The second configuration is cyclic. The mask is mosaicked in a 2×2 array and only one of the four copies of the basic pattern covers the detector completely, while the other three are free to miss it (Figure 2.16c). From the shaded areas of Figure 2.18b one can write:

$$\frac{FoV}{a} = \frac{d_d}{b} \tag{2.78}$$

hence:

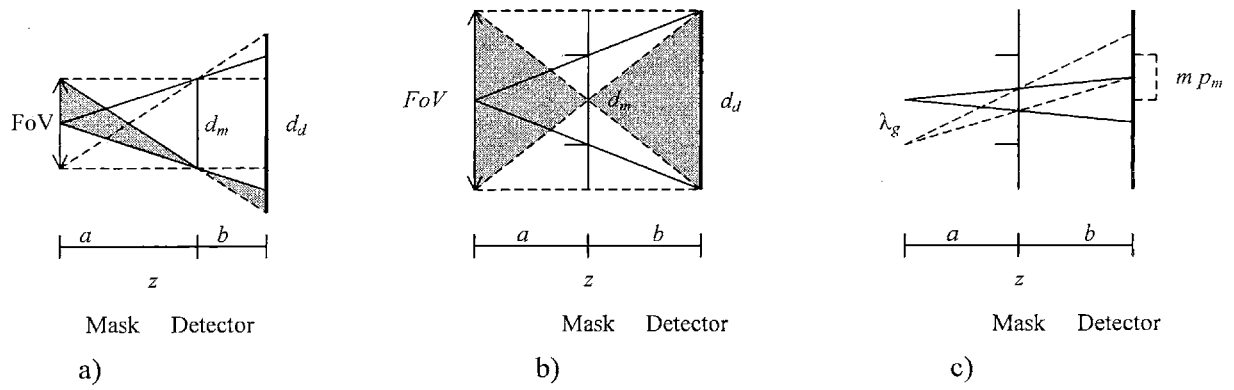


Figure 2.18: calculation of field of view for a) non-cyclic geometry and b) cyclic geometry. The mask was mosaicked by placing an elementary period at the center and two half elementary periods at the sides. All points in the field of view project a complete mask shift on the detector. c) the calculation of the geometric resolution λ_g.

$$FoV = \frac{d_d}{m-1} \tag{2.79}$$

which is readily seen to be larger than the field of view of the non-cyclic arrangement for a given detector. In both cases the field of view diverges to infinity in the far-field limit. In fact, as $a \rightarrow \infty$, $m \rightarrow 1$. In this case, however, one is interested in angular rather than absolute distance. The angular field of view, is defined as:

$$\Delta\vartheta = \arctan\left(\frac{FoV}{2a}\right) \tag{2.80}$$

The coded aperture camera can reconstruct objects perfectly within an angle $\pm\Delta\vartheta$ from its axis. In the case of a cyclic system, this is:

$$\Delta\vartheta = \arctan\left(\frac{d_d}{2b}\right) \tag{2.81}$$

Resolution can be defined in a number of ways. The most commonly accepted rule is that the resolution of an imager is the distance that must separate two point sources in the object so that their image is still perceived as two separate points. The question is, now, to give a more rigorous definition of this perception. Very often the two sources are declared separated if they are a Full-Width-at-Half-Maximum (FWHM) of the PSF apart. We adopted this definition. Note that a low value of the resolution

is good resolution. This should not be confused with the expression "high-resolution" of common use, which also indicates good resolution.

The question is reduced to the calculation of the PSF. Following the formalism of the argument of §2.4.7 a reconstruction of a point source can be obtained, but the argument must be modified because in the calculation of geometric resolution sampling effects must not be considered. Also, following the idealization of an ideal imaging system, decoding has to be performed with a continuous convolution. A suitable decoding function can be defined similarly to the function A_δ:

$$G_\delta(\vec{r}) = \sum_{k,l} \mathbf{G}_{k,l}\delta(\vec{r} - \vec{r}_{k,l}) \tag{2.82}$$

G_δ is a collection of infinitely narrow peaks displaced on the lattice $\vec{r}_{k,l}$. The projection of a point source on an ideal detector through a mask **A** with holes of shape S_m is given by eq. (2.65), which is re-written here in a form equivalent but more convenient for the following:

$$R(\vec{r}) = A(\vec{r}) = \iint_{\vec{\xi}} A'_\delta(\vec{r} - \vec{\xi})S'_m(\vec{\xi})d^2\vec{\xi} \tag{2.83}$$

With these definitions, the decoded image is now a 2d continuous function:

$$\hat{O}(\vec{r}_r) = R(\vec{r}) \otimes G_\delta(\vec{r}) = \iint_{\vec{r}} \iint_{\vec{\xi}} A'_\delta(\vec{r} - \vec{\xi})S'_m(\vec{\xi})d^2\vec{\xi}G_\delta(\vec{r} \oplus \vec{r}_r)d^2\vec{r} = \iint_{\vec{\xi}} S'_m(\vec{\xi}) \iint_{\vec{r}} A'_\delta(\vec{r} - \vec{\xi})G_\delta(\vec{r} \oplus \vec{r}_r)d^2\vec{r}d^2\vec{\xi} \tag{2.84}$$

The correlation can be shown to be $\delta(\vec{\xi} \oplus \vec{r}_r)$ (see Appendix A.6) because (**A**, **G**) is a perfect pair. Recalling that this is the reconstruction of a point source, $\hat{O} = PSF$:

$$PSF(\vec{r}_r) = \iint_{\vec{\xi}} S'_m(\vec{\xi})\delta(\vec{\xi} \oplus \vec{r}_r)d^2\vec{\xi} = S'_m(\vec{r}_r) \tag{2.85}$$

So the *PSF* has the same shape of the mask hole. To find its size we first need to revert to S_m:

$$S'_m(\vec{r}_r) = S_m(\frac{z}{a}\vec{r}_r) = S_m(m\vec{r}_r) \tag{2.86}$$

If we do not allow any modulation of the mask transparency, but, as we have done so far, just either complete transparency or opacity, the FWHM is the same as the size of the hole magnified by the factor m. This, however, is in terms of the variable $\vec{r}_r$, which covers the detector space and must be translated in terms of object space. From the definition, the question is how far two sources must be in object space for their projections to be a projection of the mask hole (which, we have just learned, is a FWHM) apart. From Figure 2.18c this spacing, which by definition is the geometric resolution of the system, is:

$$\lambda_g = \frac{m\, p_m}{b} \cdot a = \frac{m}{m-1} p_m \tag{2.87}$$

where p_m is the size of a mask hole (or pixel).

This whole argument is long but explains the logic of all factors involved in the final expression. The first is the hole size, because it determines the shape of the PSF. When projected on the detector, holes are magnified by a factor m. The magnified projection has then to be rescaled in terms of the object space, which brings about the division by the factor m-1. The net effect is that the mask hole size is multiplied by m / (m-1). Since $m \geq 1$, this function is always decreasing with m and greater than 1, so it causes a degradation in resolution, whose theoretical limit is, then, the size of the hole, obtained for $m \to \infty$, i.e. for $a \to 0$ or $b \to \infty$. Interestingly enough, in this case $FoV \to 0$ and the detector must be made infinitely large. Furthermore, the function reaches 1 asymptotically, so even big increases of m for, say, $m > 4$, bear little improvement. A better strategy is to reduce the size of the holes: in the limit, $p_m = 0$, so $\lambda_g = 0$, which is the only way to attain perfect resolution.

A relation between field of view and resolution can now be derived for the cyclic configuration. Taking the ratio of eq. (2.79) and (2.87):

$$\frac{FoV}{\lambda_g} = \frac{d_d}{m\, p_m} \tag{2.88}$$

Since the projection of the mask takes the whole detector, for all shifts:

$$d_d = m\, d_m \tag{2.89}$$

which leads to:

$$\frac{FoV}{\lambda_g} = \frac{d_m}{p_m} \tag{2.90}$$

The ratio of the mask size to the mask pixel size is the number of pixels in the side of the mask pattern n. We thus obtain the fundamental relationship:

$$FoV = \lambda_g n \tag{2.91}$$

which is independent of magnification. For a given field of view, good resolution requires the highest possible number of holes. Conversely, for a given resolution, a large field of view is possible only for large patterns. A third implication is that for a given mask, field of view can be traded for resolution. In fact, for a given mask and detector, magnification can still be changed. From eq. (2.79) a larger magnification implies a smaller field of view but, as the ratio to resolution is constant, resolution benefits (see also eq. (2.87)). Magnification can be adjusted by modifying the mask-to-detector or mask-to-object distance, or both, which can be done at constant object-to-detector distance by simply moving the mask. Note, however, that if both mask and detector are given, in a cyclic geometry, m is fixed through eq. (2.89). Changing the magnification implies not using the whole detector ($m < d_d / d_m$), which means that m is not as big as it could be. Since the active area of the detector is $m \times d_m$, this expression can be substituted in place of d_d in eq. (2.79). The field of view becomes:

$$FoV = \frac{m\, d_m}{m-1} \tag{2.92}$$

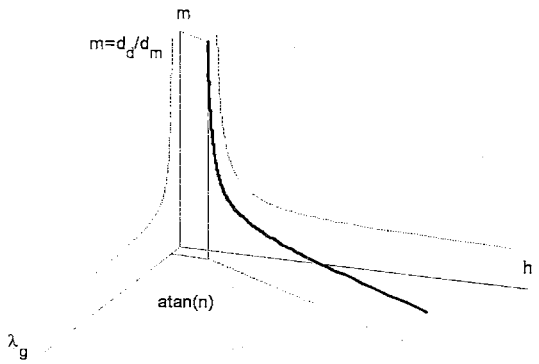


Figure 2.19: trade-off between resolution and field of view (h) for a given mask and detector.

which shows that the field of view is a decreasing function of magnification. From this and eq. (2.87), setting m according to eq. (2.89) is recognized to be a maximum-resolution minimum-field-of-view setting. A graphical visualization of the problem is provided in Figure 2.19. The lower limit for m, which is 1, does not set a limit on the maximum field of view. However, we shall see that another limit, set by detector sampling, may limit the minimum m and, thus, the maximum field of view. On the other hand, the limit on resolution is set by the maximum m, which comes from eq. (2.89).

Of course, the detector size poses limitations. The one on the field of view is clear from eq. (2.79). The advantage of having a large detector in terms of resolution is not so immediate. Solving eq. (2.79) for m and substituting in eq. (2.87) gives:

$$\lambda_g = p_m(1+\frac{FoV}{d_d}) \tag{2.93}$$

which shows that a large detector allows better resolution at constant field of view.

2.7 Decoding techniques

The issue of decoding has been long debated. In particular, attention has focused on two problems: the appropriate scaling of decoding coefficients and actually realizing a periodic correlation. To understand the origin of these problems we have to bear in mind that in the early days of coded aperture imaging optical methods were used in decoding, which limited decoding coefficients to positive values. Following the historical developments helps in understanding the solutions that were devised and, thus, current methods. The use of optical decoding techniques is obsolete for a number of reasons: non-linearity of the film, practicality, ease of data storage, difficulty to use negative and variously scaled decoding coefficients.

2.7.1 Early methods

Early authors dealing with random arrays used matched decoding, in which **G** is the same as **A**, i.e. is an array of 1s and 0s, and a non-cyclic geometry ([13]), probably because a matched filter was assumed optimal ([6]). The decoding pattern is the pattern itself, its elements being 1s and 0s. Neither array is mosaicked. A complete mask array is projected on the detector, which is not completely covered (Figure 2.16b). Typical PSFs looks like a spike on top of a pyramid whose base is as large as the projection (Figure 2.20. See also [13], [50]). The ratio between the spike and the top of the pyramid is

equal to the open fraction and is the peak-to-average-sidelobe discussed in §2.4.1. The pyramid appears because as the decoding array is shifted, the overlap between **A** and **G** decreases until, at the border, the disengagement of the arrays is complete and zero reached. Brown introduced a first solution, mismatched decoding ([50]), in which the 0s of **G** are substituted with –1s. Since the optical methods used at the time prevented the use of negative decoding coefficients, two positive decoding arrays were used and results subtracted. The average of the sidelobe level drops immediately to zero, so that the pyramid is canceled. The peak value remains the same because for no shift all –1s are canceled by opaque positions of **A**. Note however, that the effect of non-periodic correlation is still evident in the sidelobe variance, which diminishes to 0 as disengagement completes. This is different from the uniform behavior seen in §2.4.1, where periodic correlation and a mosaicked array (Figure 2.16c), a case to our knowledge not found in literature, were assumed.

Mismatched decoding works only for 50% open arrays, but can be extended to other open fractions. The result is balanced decoding ([13]): the decoding array is still assumed two-valued: be these values g_+ and g_-. For a random array of open fraction ρ the number of holes is ρN_T and the number of holes still superimposed after shift is $\sim \rho^2 N_T$. When decoding any off-peak position, these $\rho^2 N_T$ holes of **A** are multiplied by g_+ while the remaining $(\rho N_T - \rho^2 N_T)$ are multiplied by g_-. The decoding coefficients of balanced decoding are obtained by setting the reconstructed peak height to N, which gives $g_+ = 1$, and g_- so that the sidelobe value is 0:

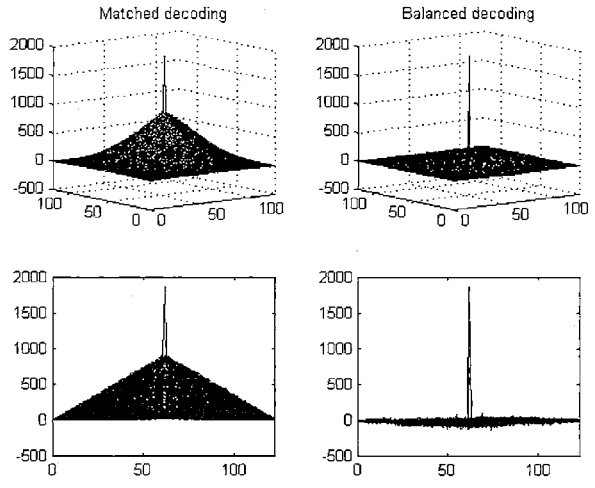


Figure 2.20: PSF of a ~50% open 62 × 62 random array for matched and mismatched (in this case the same as balanced) decoding. The bottom figures are a line-by-line plot of the 3d surfaces above them.

$$g_+ \rho^2 N_T + g_- (\rho N_T - \rho^2 N_T) = 0 \tag{2.94}$$

which gives:

$$g_- = \frac{\rho}{\rho - 1} \tag{2.95}$$

2.7.2 Periodic projection

Non-periodic correlation is used only in the decoding of random and non-redundant arrays. If arrays somehow based on cyclic difference sets are used, periodic correlation must be used. The most straightforward case is that of Figure 2.16c, where the recorded pattern is some periodic shift of the basic pattern. Decoding prescribes that the recorded projection be correlated with the decoding array. Correlations are equivalent to convolutions, provided that one of the arguments is first reflected. In their turn, correlations can be calculated via the convolution theorem, i.e. as inverse-transform of the product of the Fourier transforms of the arguments, with the relative computational advantages of Fast Fourier Transform (FFT) algorithms. Modern computers have made worries on computational time and the need of storing large decoding arrays obsolete ([6], [27]) and modern algorithms also have eliminated the annoying restriction on the size of the arguments, which used to be limited to powers of 2. Conveniently for us, the FFT algorithm assumes periodic functions so that the recorded data can be directly correlated with the decoding array with no need of zero padding or mosaicking of the decoding array. An exception is that of 1d sequences folded in rows (see §2.4.3), for which the decoding array does not have the same size as the elementary basic pattern. In this case zero padding of the data is needed. Decoding is then performed by taking the valid part (the part where superposition of the arguments is complete) of a non-periodic correlation, which is equivalent to a periodic correlation.

If the projection of a basic pattern takes less than the whole detector area because maximum magnification is not used (see section 2.6) or because of depth dependence (see section 2.8), it is easy to use data from an area around the detector center only and cut the rest, which is the same as making the detector artificially smaller to match the projection of an elementary pattern. If the decoding pattern is scaled appropriately, the PSF is not affected. Zero padding or mosaicking are not necessary.

Under these conditions there is no need to worry about the PSF smoothly connecting to 0 at the boundaries. In fact, inside the reconstruction area, which is the only area displayed, the PSF is a δ function on the top of some baseline level, which can be subtracted. In Chapter 4 is shown that all choices of the decoding coefficients related by a linear transformation are completely equivalent and do not affect

the SNR. The decoding array can then be rescaled to satisfy some criterion other than an ideal PSF, which is always the case. For $g_+ = 1$ and $g_- = 0$ matched decoding is found, for $g_+ = 1$ and $g_- = -1$ mismatched decoding is obtained and $g_+ = 1$ and $g_- = \rho / (\rho-1)$ corresponds to balanced decoding. However, other criteria may guide the choice. For example, if the number of counts in the image must be the same as the number of counts collected, the correct coefficients are $g_+ = 1$ and $g_- = (1-1/\rho N_T)\, \rho / (\rho-1)$. Other possibilities are setting the expectation value of uniform background to 0, the baseline generated by any source to 0, the peak height over its own baseline to a multiple of the peak height, the pixels in the reconstruction to be statistically independent ([51]). It turns out that combinations of these conditions can be satisfied at the same time. For example, if the peak height can be fixed, at the same time pedestals can be set to 0 and total counts conserved. This leads to mismatched decoding. On the other hand, if one wants to set peak height to some multiple and background to zero, balanced decoding is obtained. These different choices matter only as long as the absolute value of data is of interest and measurements or fits with theoretical models need to be done. The appearance of an image is independent of any of these choices.

All our experiments were carried out in periodic geometry with balanced decoding.

2.7.3 Non-periodic projection

When data are acquired in the geometry of Figure 2.16b particular care must be taken because the projection of the pattern is now surrounded by zeros (for a point source). Of course, cutting an area around the detector center as in the previous paragraph would not work because parts of the only projected period of the pattern would be lost. A possible strategy is to cut an area equal to (almost) twice the size of the projection of a mask pattern and then mosaic the decoding array in a 2×2 array (minus a row and a column to avoid aliasing, see section 2.5). Decoding with periodic or non-periodic correlation maintains an ideal PSF.

2.7.4 Detector sampling

When using digital decoding a number of choices must be made, first of all how to sample the projection[9]. So far, we have implicitly assumed that the projection of each mask hole is sampled once. More in general, the projection of each hole can be sampled on an $\alpha \times \alpha$ square of square pixels. This leads to another fundamental equation:

$$m\,p_m = \alpha\,p_d \tag{2.96}$$

where p_d is the dimension of a detector pixel. This equation represents the passage from the continuous to the digital representation. The shadow of size $m\,p_m$ is the specialization to the case of square holes of S'_m, while one of the $\alpha \times \alpha$ pixels is a similar specialization of S_p. If the mask is a $r \times s$ array, the data is recorded on a $\alpha r \times \alpha s$ array where each $\alpha \times \alpha$ square is filled with 1s or 0s, according to the mask. It is surely possible to collapse the array back into an $r \times s$ array by summing over the $\alpha \times \alpha$ squares, but this would simply bring us back to the previous case. For obvious reasons, the case $\alpha > 1$ was given the name "fine sampling" ([44]).

The decoding array also has to be extended from $r \times s$ to $\alpha r \times \alpha s$. Since we still have to match the mask shape, each location in the original decoding array is expanded to a $\alpha \times \alpha$ square, whose new locations must then be filled. This can be done in at least two ways: (i) by using the same coefficients of the original decoding array; or (ii) 0s. The latter method is called δ decoding ([44]). It is interesting to analyze the PSF obtained in these two cases.

The analysis starts once more from the projection of a point source. In this case the form factor is (from eq. (2.69)):

$$\mathbf{F}_{r-i,s-j} = \mathbf{H}\left(\frac{r-i}{\alpha-1}, \frac{s-j}{\alpha-1}\right) \tag{2.97}$$

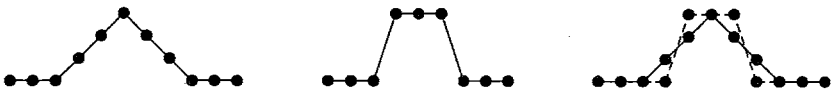
where

$$\mathbf{H}(i,j) = \begin{cases} 1 & 0 \le i,j \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.98}$$

i.e. an $\alpha \times \alpha$ square. Note that this time the grid over which $\mathbf{F}$ is defined is finer than the grid over which $\mathbf{A}_{i,j}$ is defined because the projection is fine-sampled. In the case of a point source, let the recorded pattern $\mathbf{R}_{r,s} = \mathbf{A}^F_{k,l}$, where $\mathbf{A}^F_{k,l}$ is the finely sampled (hence the superscript F) array obtained from $\mathbf{A}_{i,j}$ by expanding each location to an $\alpha \times \alpha$ square and then filling the new positions with 0s and 1s according to the original element. It is also convenient to define the array $\mathbf{A}^{F\delta}_{k,l}$. As $\mathbf{A}^F_{k,l}$, it is defined as an expansion

---

[9] In the following, misalignment between mask projection and detector sampling, that will be analyzed in Chapter 5, is neglected. The projection of a mask hole is assumed to fall exactly on some number of detector pixels.



a) b) c)

Figure 2.21: comparison of the PSF for a) fine sampling and b) fine sampling and δ decoding. In this case $\alpha = 3$.

of $\mathbf{A}_{i,j}$. This time the $\alpha \times \alpha$ squares are filled, in the new positions, by 0s (hence the superscript δ, by analogy with δ decoding). Clearly:

$$\mathbf{A}_{k,l}^{F} = \mathbf{A}_{k,l}^{F\delta} * \mathbf{H} \tag{2.99}$$

Similar definitions can be given for $\mathbf{G}_{k,l}^{F}$ and $\mathbf{G}_{k,l}^{F\delta}$. It is important to recognize that

$$\mathbf{A}_{k,l}^{F\delta} \otimes \mathbf{G}_{k,l}^{F\delta} = \delta \quad \text{but} \quad \mathbf{A}_{k,l}^{F} \otimes \mathbf{G}_{k,l}^{F} \neq \delta \tag{2.100}$$

In fact, in the first case, when the shift is such that original positions of $\mathbf{A}_{k,l}^{F\delta}$ are superimposed to original positions of $\mathbf{G}_{k,l}^{F\delta}$, the point by point multiplication is zero at all new points, so the value of the correlation is exactly the same as for $\mathbf{A} \otimes \mathbf{G} = \delta$. For all other shifts, original positions of $\mathbf{A}_{k,l}^{F\delta}$ are superimposed to 0s of $\mathbf{G}_{k,l}^{F\delta}$, so that the correlation is once again 0 outside the peak. On the other hand, from their definition:

$$\mathbf{A}^{F} \otimes \mathbf{G}^{F} = (\mathbf{A}^{F\delta} * \mathbf{H}) \otimes (\mathbf{G}^{F\delta} * \mathbf{H}) = \Re[(\mathbf{G}^{F\delta} \otimes \mathbf{A}^{F\delta}) * \mathbf{H}] * \mathbf{H} \tag{2.101}$$

where $\Re$ is the reflection operator and the result of Appendix A.5 was applied twice and that of Appendix A.4 once. Since $\mathbf{A}_{k,l}^{F\delta} \otimes \mathbf{G}_{k,l}^{F\delta} = \delta$, applying Appendix A.4 once more gives:

$$\mathbf{A}^{F} \otimes \mathbf{G}^{F} = \Re[\delta * \mathbf{H}] * \mathbf{H} = \Re[\mathbf{H}] * \mathbf{H} = \mathbf{H} \times \mathbf{H} \tag{2.102}$$

If δ decoding is used the PSF becomes:

$$PSF = \mathbf{A}^{F} \otimes \mathbf{G}^{F\delta} = \Re[\mathbf{G}^{F\delta} \otimes (\mathbf{A}^{F\delta} * \mathbf{H})] = \Re[(\mathbf{G}^{F\delta} \otimes \mathbf{A}^{F\delta}) * \mathbf{H}] = \Re[\mathbf{H}] \tag{2.103}$$

In neither case is the PSF a δ function: with fine sampling the details of the finite dimension of the holes are now seen. The two PSFs are compared in Figure 2.21. They are seen to have the same FWHM, which is the same resolution, according to our definition. Still, the FWHM does not tell the whole story. Possibly because of the sharper peak, δ decoded images seem to have somewhat better resolution ([44]). Note, however, that in δ decoding each position is reconstructed starting from fewer data points than in normal decoding. Actually, from the two PSFs it is clear that normal decoding is equivalent to δ decoding followed by convolution with **H**. This is an averaging operation over $\alpha \times \alpha$ pixels, which is an example of low-pass filtering. In conclusion, δ decoding provides better resolution but also worse statistics. It is convenient because the result of normal decoding is just a convolution away.

2.8 Depth of focus and 3d laminography

From the discussion of section 2.7 it is clear that the decoding array must be scaled to match the dimensions of the projected pattern which is, by definition of *m*:

$$m\,n\,p_m = \left(1+\frac{b}{a}\right) n\,p_m \tag{2.104}$$

where *n* is the number of pixels in the mask (side only). If the match is less than perfect, the PSF is corrupted. When a thick object is imaged, all object planes parallel to the detector have different object-to-detector distances *a* and each plane is associated to a different size of the projection. Only one depth at a time can be decoded with ideal PSF; all others still contribute to the image because they are still present in the recorded data, but their reconstruction is corrupted by a non-ideal PSF. From an *a-priori* point of view, this can be bad, because distorted signals appear in the image, as well as good, because if all other planes were blurred in a uniform, featureless background, the only plane with ideal PSF would stand out sharply. Unfortunately it is difficult to predict which is the case, but the potential for 3d laminography is much advertised in literature ([13], [6]) and has been investigated experimentally ([33]).

The best indicator of performance is again the PSF, this time as a function of three variables, the most interesting being depth. This investigation must be carried out either experimentally or by simulation. Results are in Chapter 3 and Chapter 8. However it is still possible and interesting to study analytically how sensitive the mask sampling is to depth variation. Substitution of eq.(2.74) in eq. (2.96) yields:

$$\alpha\, p_d = \left(1+\frac{b}{a}\right) p_m \tag{2.105}$$

An experiment is performed with a given mask, detector and a given mask-to-detector distance. Keeping track of constants, differentiation of both sides gives:

$$\frac{d\alpha}{da} = -\frac{p_m b}{p_d a^2} \tag{2.106}$$

This derivative shows the sensitivity of the size of the projection (in units of detector pixels) to changes in object-to-detector distance. The negative sign implies that as the object nears the mask, the size of the mask projection increases. The dependence of the projection size on a is greater for small a, i.e. for high magnification. In this case, we expect a small depth of focus because planes not too far from the focal plane are already out of focus. A 3d PSF is obtained repeating the decoding of the projection of a point source for several depths. Depth of focus can be defined as the FWHM of this PSF measured along this depth variable. Experiments and simulations will show that the FWHM along depth is about one order of magnitude worse than the FWHM along transverse directions.

Eliminating α with eq. (2.96), we can assess the variation per image pixel ($d\alpha / \alpha$) due to a displacement da:

$$\frac{d\alpha}{\alpha\, da} = -\frac{b}{za} = -\frac{m-1}{z} \tag{2.107}$$

A first observation is that maximum defocusing is obtained at low object-to-detector distance and for high magnification. This suggests to image as close as possible to the detector, near-field artifacts allowing ([33]). The final design parameters will turn out to be $z = 40$ cm, $a \cong 10$ cm, $b \cong 30$ cm, so $d\alpha / (\alpha\, da) \cong 0.075$ / cm, i.e. the size of the pattern changes by 7.5% for a displacement of 1 cm from the focal plane. A second observation is that this equation is useful in fine focusing when experimental placement uncertainties need to be accommodated. An adjustment procedure is as follows: take an image of a point source; measure on the screen how many detector pixels the projection of a mask period takes; compare with the design parameter; calculate $d\alpha$ as the design α minus the measured α; use the measured values in:

$$da = -\frac{za}{\alpha b} d\alpha \tag{2.108}$$

to obtain an estimate of how much the mask should be moved.

Dependence of m on depth leads to different resolution and field of view for the different planes. A limit is set on the plane with maximum magnification, i.e. the one closest to the mask. Since an entire elementary pattern must cover the detector, $m \le d_d / d_m$. Substitution of eq.(2.74) gives:

$$a \ge \frac{b}{\frac{d_d}{d_m} - 1} \tag{2.109}$$

Planes closer than a to the mask can not be reconstructed with an ideal PSF because they cast too big a projection.

PART II:

METHODS, THEORETICAL ADVANCEMENTS AND EXPERIMENTAL RESULTS

Chapter 3 SIMULATION TOOLS: COMPUTER CODES AND OPTICAL BENCH

The design of a mask involves many parameters many of which can not be optimized in any other way than trial and error. Some others are determined using new theoretical findings involving approximations that suggest some kind of verification prior to finalization of the design. In this Chapter the two simulation tools used are analyzed. The first is an original computer code. It is not a Monte-Carlo method in that it does not attempt at simulating the physics of the process particle by particle. Rather, projections are calculated according to geometric optics. Of course, a major advantage of computer simulation is the ability to switch pieces of physics on and off as needed to gain insight.

The second simulation tool is an optical analogue of the coded aperture camera. The experiment uses visible light and a CCD camera in place of γ-rays and an Anger camera. It is set up on an optical bench placed in a light-tight box. This experiment provided experience with problems likely to be found in a real environment, where alignment, placement of components and size of objects can not be as regular and ideal as in a fast computer simulation. Indeed, a strong deviation from isotropy of a semi-transparent screen greatly enhanced near-field effects, leading to the realization that, far from being a bug in simulation codes, they would be a primary issue in system design.

In this Chapter are described the two simulators and some of the results they produced. The predictions of the computer code are compared to results from the optical simulator and data obtained from a Siemens E-Cam in a preliminary experiment. With the goal to provide a better understanding of the system the codes intend to simulate, the first section gives a very brief description of the principles of operation of an Anger camera.

3.1 The Siemens E-Cam

An Anger camera is a single, large (53.3 × 38.7 cm active area for the Siemens E-Cam), crystal of scintillating material (0.95-cm-thick NaI(Tl)) to which are optically coupled about 60 photo-multiplying tubes which convert a light signal to an electric pulse. When a γ-ray interacts in the crystal, light is produced at the point of interaction as well as in its neighborhood and is emitted in all directions. Each

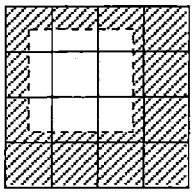
phototube is illuminated differently, depending on the location of the event, whose center of mass is calculated by processing the signal from different tubes. Events do not look like a dimensionless point, but are spread over a region. This blur due to the physics of the process is measured by the intrinsic PSF of the detector. Technical data for the Siemens E-Cam report for the intrinsic PSF a 3.7-mm FHWM at the center of the crystal for 140 keV photons, increasing to 3.9 mm at the edges of the crystal.

The position of the event is then located on a square grid of pixels and a counter associated to the corresponding pixel increased. The final image is this distribution of number of events over space. The details of the discretization are treated at length in Chapter 7.

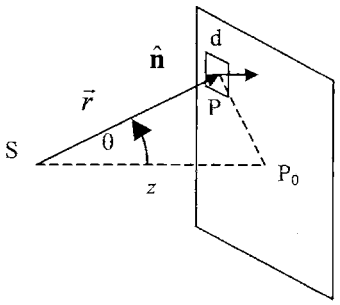
3.2 The simulation code

The core of the simulation code is a module that calculates where the projection of a single mask hole cast by a point source falls on the pixel grid of the detector. The code is reported in Appendix D.2. Mask holes are assumed square. Applying purely geometric formulae, the code finds the location of the borders of the projection of the hole by ray tracing. For a given point source and a given mask hole, a detector pixel can be totally, partially or not illuminated at all (the vast majority of pixels). In the first case a 1 is stored in the pixel, in the last a 0. Partially illuminated pixels are assigned a number equal to the illuminated fraction of their area. An example is given in Figure 3.1a-b. The process is then repeated for all mask holes. To accelerate the procedure, mask positions are not considered one at a time, but are clustered in larger squares when they form one (see Appendix D.1). This greatly accelerates the simulation of URA patterns, while it is of no benefit for NTHT arrays which, however, have a relatively small number of holes anyway. When the projection of a mask pattern is completed, a correction is made to include the effect of mask transparency. The correction varies over the mask even for masks of constant thickness because of the different incidence angle of γ-rays at the mask, oblique rays effectively seeing a thicker mask. All detector pixels are then multiplied by the activity of the point source, the exposure time and the solid angle subtended by the central pixel. Each pixel is then multiplied by a spatially dependent factor, the cubed cosine of the incidence angle at the pixel. This factor comes from the inverse square law and an obliquity factor. With reference to Figure 3.1c, the number of photons passing through an infinitesimal area dA around point P and originating at S is:

$$I(\mathrm{P})\,dA = \frac{I(\mathrm{S})}{4\pi r^2}\hat{\mathbf{r}} \circ \hat{\mathbf{n}}\,dA \tag{3.1}$$



0.04	0.2	0.16	0
0.2	1	0.8	0
0.14	0.7	0.56	0
0	0	0	0



a) b) c)

Figure 3.1: a) projection of a mask hole (dashed line) over a 4 × 4 area of the detector b) values applicable to the pixels of a). No mask transparency assumed. c) Definition of terms for the evaluation of the flux at a pixel.

where the caret indicates unit vectors. Since $z = r\cos(\vartheta)$ and $\hat{\mathbf{r}} \circ \hat{\mathbf{n}} = \cos(\vartheta)$:

$$I(P)\,dA = \frac{I(\mathrm{S})}{4\pi z^2}\cos^3(\vartheta)\,dA \tag{3.2}$$

At the end of these multiplications each pixel contains the average number of counts due to a point source in the object. The result is stored and is later added to the results from all other point sources. The point sources are placed at the knots of a 3d orthogonal grid, so that thick objects can be simulated. The grid pitch can be adjusted independently in the three dimensions. Continuous objects were simulated by setting the grid pitch to a value smaller than the design resolution of the coded aperture camera, compatibly with execution time. The thyroid shape extensively used in our simulations is a 91 × 79 grid with a pitch of 0.9 mm (Figure 7.12a). 3820 points have some activity. All of them must be projected through all mask holes. For a NTHT mask with 480 holes, projection takes about 2 hours on a 333MHz Pentium II personal computer. The code was run as a non-compiled Matlab script.

This basic procedure handles the effect of discrete pixels and near-field geometry. However, many other non-idealities are left out: finite mask thickness, the intrinsic PSF of the detector, statistical noise and overflow. The final projection is saved in a dedicated variable before these other effects are added to avoid a whole recalculation when only some parameters not influencing geometry, such as intrinsic PSF, exposure times and activity, change. The effect of PSF blurring and statistical noise are added in a few seconds.

3.2.1 Finite mask thickness

Simulation of mask thickness is fundamental for the optimization of this dimension and the object-to-detector distance. The code is fairly complex and is best explained following the history of its development.

The mask is treated as made of a variable number of layers, to be optimized on the basis of a trade-off between accuracy and execution time. In our cases a value of 7 seemed best, which leads to simulation times of 14 hours for the thyroid phantom. The question is how to combine projections of different layers. In early versions of the code, the projection of the whole object through each layer, supposed isolated, was calculated first. Since multiplication by the source irradiance was delayed to the last layer, each layer stored values between 0 and 1. Following the logic that a ray must pass through all layers to reach the detector, projections were multiplied point by point. This procedure is wrong because it holds for the same point sources only. To see this with an example, assume a uniform object. The projection is (see Appendix A.2):

$$\mathbf{R} = \mathbf{O} \times \mathbf{A} = \text{constant} \tag{3.3}$$

for every layer. Point by point multiplication of the results from different layers simply changes this constant, uniformly over space. On the other hand, if the projection of a point source through a whole layer is calculated and then multiplied point by point with the projection of the same point source through a second layer, the multiplication has the effect of canceling from the first layer the fraction of rays that do not pass through the second layer. For example, if a pixel is completely illuminated through a layer, but only partially through another, the product would substitute the 1 assigned by the first projection with the fraction of rays passed by the second layer. This is why it is important that the cycle over mask slices be carried out inside the cycle over point sources.

Unfortunately this approach still has a serious problem. In fact, if only half of a certain pixel is illuminated through a certain layer and if incidence angles are not very high, the same pixel is also illuminated in a very similar way through an adjacent layer, i.e. the illuminated region does not change much. Yet, if the product of the illuminations were taken, the result would be that only a fourth of the pixel is illuminated. The problem is that since adjacent layers look very much similar to impinging γ-rays there is a strong correlation between the probability of passing through a layer and the successive. So the probability of passing through both layers is not the product of the probabilities of passing through the individual layers. This explained why the codes were sensibly underestimating the total number of counts when thick masks were simulated.

The problem was solved by defining an equivalent illumination fraction. Be f and f' the illumination fractions of a detector pixel due to the first and the next adjacent layer when taken individually. The equivalent illumination fraction of the first layer is the sum of the fractions of rays that reach directly the detector plus the fraction of those that penetrate the mask:

$$(1-f)e^{-\mu\rho t_m/\cos(\alpha)}+f \tag{3.4}$$

where μ is the attenuation coefficient, ρ the density of the material, α the incidence angle and t_m the thickness of a mask layer. In the hypothesis that the shadow cast by one of the two layers completely covers the shadow cast by the other, the combined equivalent fraction allowed on the pixel is:

$$\begin{cases}(1-f')e^{-2\mu\rho t_m/\cos(\alpha)}+(f'-f)e^{-\mu\rho t_m/\cos(\alpha)}+f & \text{for } f<f' \\ (1-f)e^{-2\mu\rho t_m/\cos(\alpha)}+(f-f')e^{-\mu\rho t_m/\cos(\alpha)}+f' & \text{for } f'<f\end{cases} \tag{3.5}$$

Defining:

$$\begin{cases}\bar{f}^{(2)}=f'+(f-f')e^{\mu\rho t_m/\cos(\alpha)} & \text{for } f<f' \\ \bar{f}^{(2)}=f+(f'-f)e^{\mu\rho t_m/\cos(\alpha)} & \text{for } f'<f\end{cases} \tag{3.6}$$

the simpler form:

$$(1-\bar{f}^{(2)})e^{-2\mu\rho t_m/\cos(\alpha)}+f^{(2)},\quad \text{with } f^{(2)}=\min(f,f') \tag{3.7}$$

is reached. The argument can be iterated for all other slices. In general:

$$\begin{cases}\bar{f}^{(i)}=f'+(f^{(i-1)}-f')e^{\mu\rho t_m/\cos(\alpha)} & \text{for } f^{(i-1)}<f' \\ \bar{f}^{(i)}=f^{(i-1)}+(f'-f^{(i-1)})e^{-(i-1)\mu\rho t_m/\cos(\alpha)} & \text{for } f^{(i-1)}>f\end{cases},\ \text{with } f^{(i)}=\min(f^{(i-1)},f') \tag{3.8}$$

where f' is the illumination fraction of the ith layer. It can be shown that with the new technique, if illumination fractions of all n slices are equal to f, the combined illumination coefficient is:

$$\bar{f}^{(n)}=(1-f)e^{-n\mu\rho t_m/\cos(\alpha)}+f \tag{3.9}$$

which is the exact result for an infinitely thin mask. Neglecting correlations as explained above, the result would be:

$$\left((1-f)e^{-\mu\rho t_m/\cos(\alpha)}+f\right)^n \tag{3.10}$$

which vanishes for $n \to \infty$. However, in both cases the limiting situations in which $f = 0$ and $f = 1$ are handled correctly.

Not even the solution in terms of the average illumination is exact because the hypothesis on which is based (that the shadow from one layer completely cover or be covered from the shadow of the next) is not always respected. However, in practical cases mask layers are very close, and incidence angles not very high, so projections of adjacent mask layers do not greatly shift, and the hypothesis is not far from reality.

The layer-by-layer approach makes it possible to simulate holes of varying cross-section with no significant different in execution times.

3.2.2 Detector PSF

Assuming a constant PSF over the detector makes it possible to convolve projection data with a 2d Gaussian of standard deviation:

$$\sigma = \frac{FWHM}{2\sqrt{2\ln 2}} = 1.57 \text{ mm} \tag{3.11}$$

to simulate intrinsic PSF blurring (FWHM = 3.7 mm). The curve was normalized to conserve the total number of counts:

$$PSF_i(\vec{r}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|\vec{r}|^2}{2\sigma^2}\right) \tag{3.12}$$

3.2.3 Statistical noise

Once the blurring is complete, each pixel stores a realistic average value of the photon count. Coming from radioactive decay, counts at all pixels are independent. Strictly speaking their distribution is binomial, but radioactive processes satisfy the conditions for the Poisson distribution to be an excellent approximation, as long as counts are taken for periods of time much shorter than the half life of the isotope involved. Sampling a Poisson distribution is cumbersome but necessary because a Gaussian approximation holds only in the case of a high average of counts (> 15). If a Gaussian approximation

were assumed, it would be possible, and indeed quite likely, to have negative values after noise is added, which is clearly not physically possible. However, when the expected value of counts was greater than 15, Gaussian statistics was adopted to accelerate simulations.

A routine not provided by most software packages was used to sample from a Poisson distribution. It is provided with an explanation of its basics in Appendix D.4.

3.2.4 Overflow

In the parts of the projection where noise was added according to the Gaussian model the integer part of the result must be taken because a real value must be integer. Note that this is not adding digitization noise, but it is merely making sure that the Gaussian approximation does not yield physically unreasonable results. 16 bit memory was also assumed, so any counts in excess of 65535 were reset to this value.

3.2.5 Decoding

After the projection is calculated and blurring and noise are added, zero-order artifact correction is applied[10] (see §5.4.3) and the image decoded (Appendix D.3). The technique of choice is δ decoding because other methods' results can be obtained with an additional convolution (see §2.7.4). Initially the code decoded only with integer values of α (for its definition see §2.7.4). Successive developments (section 6.3), eliminated this limitation.

3.2.6 Some sample simulations

In Figure 3.2 are provided some sample PSFs of the coded aperture camera. The projection of a 62×62 NTHT array (13.81 cm side) on 124×124 pixels (29.74 cm) of a larger detector placed at 4 m was simulated with different conditions. Note that $\alpha = 2$. When no noise or blurring is present and the mask is infinitely thin, the PSF is a rectangle, two pixel wide, as expected for δ decoding. The effect of the intrinsic PSF (FWHM = 3.7 mm, i.e. 1.54 detector pixels) is to spread the rectangle over first neighbors and, thus, also lower the peak height. Noise was then added by lowering the activity of the source. The peak is now uneven and sidelobes are not flat. Finally, the activity was restored to the previous value and a 0.7-mm-thick mask was used in place of an infinitely thin (but still attenuating)

[10] This correction is not applied to the results of the rest of this Chapter, unless otherwise stated, because artifact theory had not yet been developed at the time.

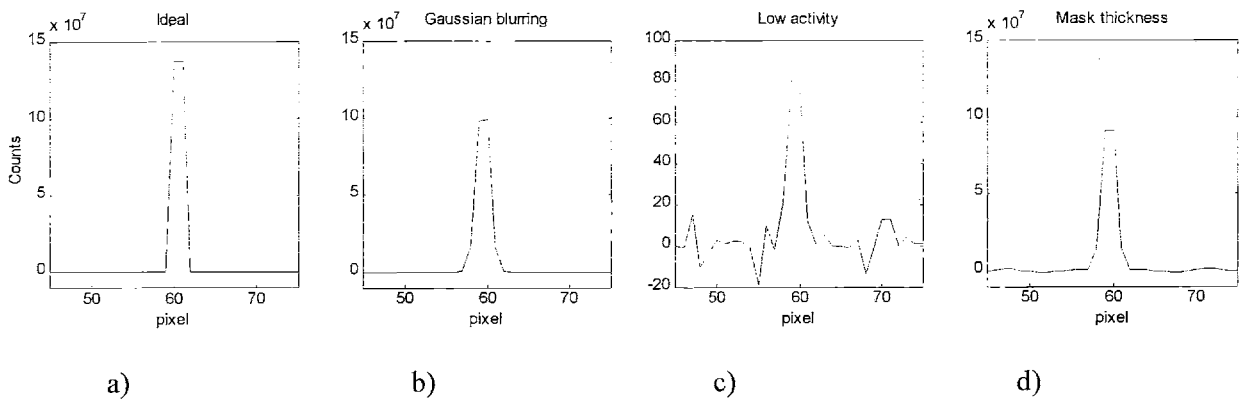


Figure 3.2: sample simulations of PSFs for a 62 × 62 NTHT MURA with a mask pixel size of 1.114 mm. a) No blurring or noise. Mask opaque but infinitely thin. b) Gaussian blurring with FWHM 1.54 pixels. c) Noise added by lowering the activity by a factor of 1 million. d) Activity restored to original value but finite mask thickness introduced. All simulations in far-field (object-to-detector 4 m).

mask. Some variations in the sidelobes are evident and show that mask thickness may be an issue. This is even more relevant at distances lower than the 4 m of this simulation, because incidence angles at the mask increase until a thickness of few millimeters becomes more significant.

The effect of finite mask thickness is twofold: thin masks project a pattern close to ideal, but are partially transparent and do not stop rays that should be blocked, allowing a higher background to corrupt the signal. Thicker masks reduce such background but also cut rays that should pass, distorting the shape

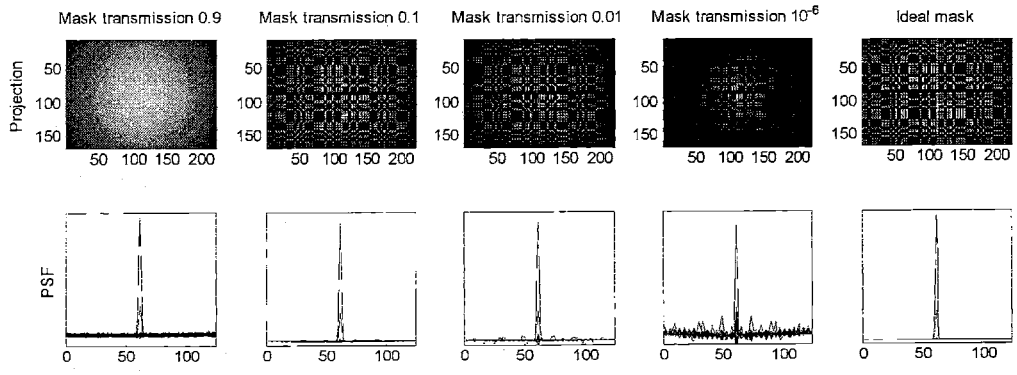


Figure 3.3: projection and PSF for increasing mask thickness. The PSFs are a column-by-column plot of a 2d distribution. The ideal mask is a 62 × 62 NTHT MURA, pixel size 1.114 mm, 1% transparent but with no thickness. A tungsten mask and ^{99m}Tc were assumed ($\mu = 1.56$ cm^2 / g at 140 keV, $\rho = 19.3$ g / cm^3). Mask thickness: 0.035 mm (transmission: 0.9), 0.763 mm (0.1), 1.527 mm (0.01), and 4.580 mm (10^{-6}). Object-to-detector distance: 40 cm. Point source: 200μCi, exposed for 1053 s. Zero-order artifact correction applied to isolate the effects of mask thickness from near-field artifacts.

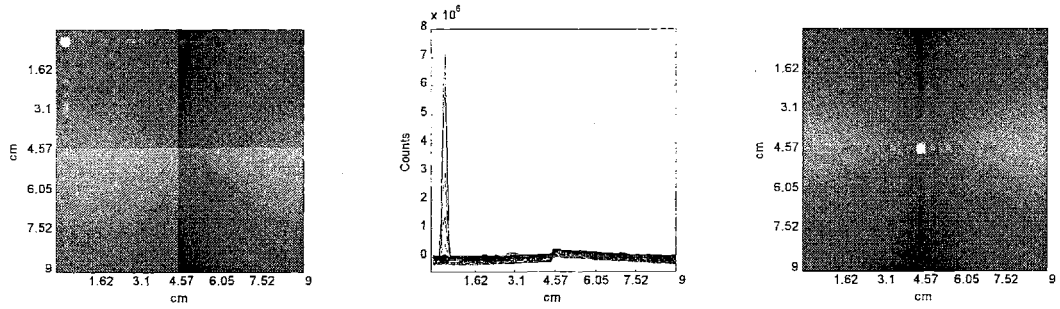


Figure 3.4: Simulations of near-field artifacts for point sources. Contrast is greatly enhanced for display purposes. The simulated mask was a non-centered MURA 61 × 61. The FoV was 9 cm. Mask 70% transparent, but infinitely thin to rule out thickness artifacts. Object-to-detector distance 40 cm. a) Point source at a corner of the FoV. b) The brightness of a) is plotted column by column to show the actual importance of the "cross" artifact. Note the discontinuity at the center of the field of view. c) Point source at the center of the FoV. Note the "bows" entering from the sides.

of the pattern and creating artifacts. This conflict explains the trends of Figure 3.3, where the effects on the PSF are also shown. Note in the sidelobes the progressive reduction of noise (due to lower transmission) and then the appearance of spurious peaks (due to pattern distortion).

At low object-to-detector distance near-field effects also become an issue. In the projections of Figure 3.3 the cosine cubed modulation is clearly visible but it can be removed with a zero order correction (§5.4.3) in the case of a point source at the center of the field of view, which we did to concentrate on the effects of thickness. When the point source is moved, results like those of Figure 3.4 are obtained. Given the obvious regularities, the phenomenon is clearly not due to random noise. While in this figure artifacts had to be emphasized with an increased contrast, it gives a good idea of what happens in the case of extended objects. In fact, even if the various point sources are spread over the field of view, they all seem to contribute to the "cross" and "bows" artifacts in the same way, so that they build up, source by source, until images are so badly affected that the object is almost lost. Given the dramatic implications, another simulation tool was needed to verify the trustworthiness of these predictions of the computer code.

3.3 The optical simulator

Designing the computer code to simulate non-idealities, such as a rotated or non-uniform mask, mask or object scatter and non-uniform unencoded background, would have slowed the calculations excessively. A physical simulation experiment was prepared to get some idea of if and how these and

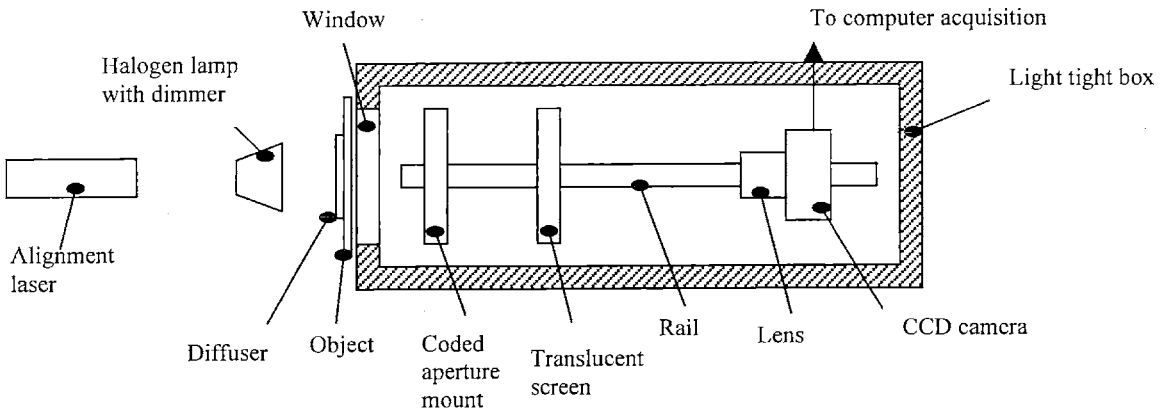


Figure 3.5: top view of the optical experiment setup.

other unexpected factors would affect images. Predicting outcomes of this optical simulator with the computer code would also be a good means of verifying the validity of computer simulations.

For practical reasons, the experiment uses incoherent visible light from a 60 W halogen flood lamp in place of γ-rays. Light is shaped into an object with a thin black cardboard sheet in which holes are pierced with a pin ("object" in Figure 3.5). The object is in contact with a piece of diffusing glass that scatters light just before it passes through these holes. This has two reasons. First, to simulate a radiation source, the points of the object must emit radiation isotropically, at least at the angles that are going to hit the detector. If there were no diffuser, photons passing through the hole would mainly have angles close to the system axis because the holes are thick enough to provide some collimation. Second, if the object is a point source, i.e. a pinhole, with no diffuser and without the coded aperture in place features of the light source, such as the pattern of the glass bulb or the filament, if an incandescence lamp is used, appear on the screen. In other words, the object would work as a pinhole camera rather than as a source. The second reason why the diffuser is used is to cancel memory of the features of the original source.

As the object is very dim because of its size and the diffuser, imaging must take place in a light-tight box. Lamp and object are placed outside it to avoid overheating. Inside the box, movable mounts for the coded aperture, a translucent screen and a CCD camera are provided by an optical rail. The coded aperture is printed directly on transparency with a laser printer. To block light missing the mask pattern, the mask is taped to a paper sheet having a square hole of the size of the mosaicked mask at its center. The sheet is in its turn taped to a glass slab that provided stiffness and mounted on the optical rail. The screen is a semitransparent diffuser glass. Its task is to intercept the rays passed through the coded aperture. The image formed on its back is digitized by a 800 × 960 pixel CCD camera focused on the

back of the screen. The combination of screen and CCD camera simulate the Anger camera. The system is aligned with the help of an external laser beam.

3.3.1 Some results

The optical simulator was used with masks under investigation for use with the Anger camera. Two apertures were considered to see the effect of different array families: the 61×61 MURA and the 63×65 m-sequence of Figure 3.6. Parameters were chosen with the aim to build a 1 to 3 copy of a realistic system. Accordingly, masks were made with a pixel size of 0.377 mm and images taken with the CCD camera were cut and pixels clustered to simulate a 1 to 3 copy of the real detector. The test object was a set of pinholes forming the letter H. To test resolution at the bottom right two pinholes were pierced as close as possible. Since the field of view of the real system is 9 cm, the letter H was made about 3-cm tall (2.8 cm) to allow a direct test of the FoV. The closest pinholes were separated by 0.7 mm, the closest distance we could obtain with a pin and cardboard from the back of a note pad. From eq. (2.91) the predicted geometrical resolution of the system is $30 \text{ mm} / 61 \cong 0.5 \text{ mm}$. In this case, it is a good approximation of the system resolution because the light experiment with a CCD camera is not affected by the 3.7-mm intrinsic PSF of γ-rays on an Anger camera. The system should be capable of resolving even the closest pinholes of the letter H.

In Figure 3.7a experimental data show the field of view and resolution are as expected: all pinholes are resolved and the letter is about the size of the reconstruction area. Also, the image is corrupted by artifacts. In Figure 3.7b is a series of simulations that, starting from ideal conditions, adds non-idealities one at a time. At the top left only near-field geometry is assumed. At the top right the mask is made 70% transparent. This explains the dark areas at the sides. Then the central row of the letter H is

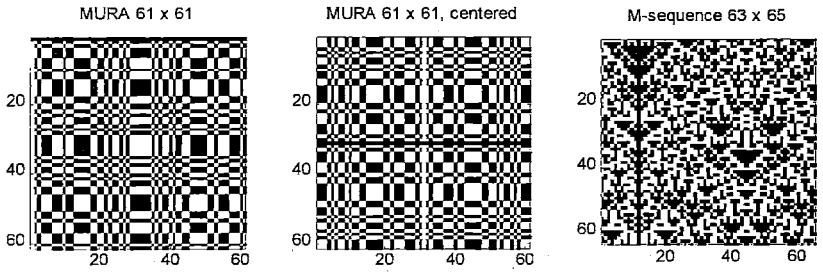


Figure 3.6: masks used for physical simulations. The mask are a 61×61 MURA and a 63×65 m-sequence.

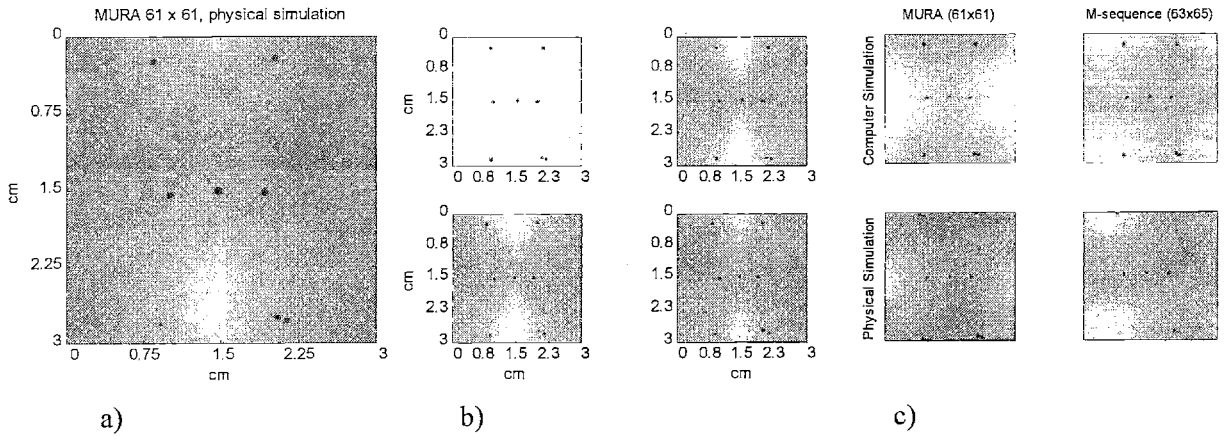


Figure 3.7: a) physical simulation with a MURA 61 × 61. b) Computer simulations progressively adding deviations from ideality provide insight in the artifacts seen in a). In all simulations the mask is partially transparent but infinitely thin. This makes sense for the optical experiment because the mask is actually a printed transparency. c) Comparison between computer prediction and physical simulation for different masks.

assumed to be 5 mm below the center of the field of view, which enhances the horizontal step of brightness seen just above the center of the H (bottom left). Finally, a slight displacement is assumed to make $\alpha = 1.98$ in place of 2. This enhances the vertical and horizontal artifacts seen to start from all point sources (bottom right).

In conclusion, the computer code seems to predict very reliably a number of effects, even for different mask families (Figure 3.7c). While for the MURA family bright areas appear at the side of the H, for the m-sequence they take the corners of the field of view.

3.3.2 Design advance

Once work with the optical simulator indicated the reliability of the computer code, design efforts were made to improve the quality of the images. The first example is mask centering.

Mask centering

An MURA pattern is basically made by squares more or less uniformly distributed. The most noticeable exception are the first row and column. This irregularity suggested that the particular shift of a mask may influence the image in near-field geometry. An image of the test object was taken with the centered MURA pattern of Figure 3.6. In Figure 3.8 the vertical line cutting across the center of Figure 3.7 has completely disappeared, while the horizontal has moved to the top (this result is explained with a small misalignment in the vertical direction). Other artifacts, however, remain. This was the first

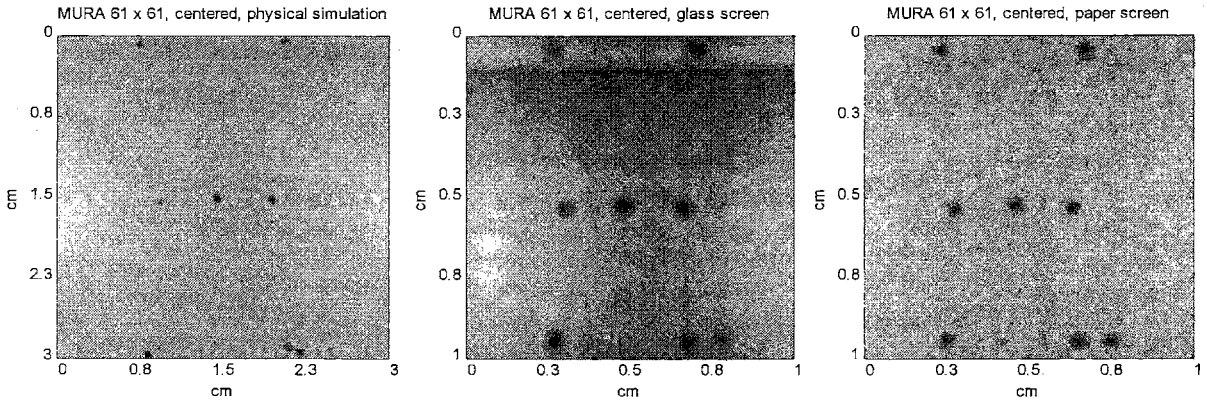
experimental evidence that near-field artifacts could be predicted by computer simulation and, if not eliminated, eased.

Screen materials

Non-idealities involved in the physical reconstruction do make things worse than initially predicted by computer simulation. The major villain is the screen. Ideally, it should transmit light uniformly. Since light coming from a point source is modulated by the cubed cosine of the incidence angle, light transmitted through the mask should follow the same distribution. Unfortunately, this is not true for the glass screen, which is found to attenuate peripheral areas much more than the center. Empirically, a distribution modulated by a cosine elevated to a power much higher than 3 is measured. The result is an apparent enhancement of near-field effects.

A way of reducing this effect is to build a smaller-scale model so that only the center of the screen, where incidence angles are still low, is used. A 1-to-9 model was built: the mask had a pixel size of 127 μm, i.e. 3 dots per pixel at 600 dpi. With this, the geometric resolution of the system is 164 μm and the FoV is reduced to 1 cm. A 1-to-3 model of the previous test object was also built. Despite the best efforts, the closest pinholes were now 1.1 mm apart.

The result is in Figure 3.8b. Near-field effects are still severe. Resolution is such that pinholes ($\varnothing = 0.63$ mm) now appear finite and spread over some pixels. Materials other than glass, such as mylar, were tried for the screen. The best result was obtained with a thin paper screen. Near-field artifacts disappear, indicating more ideal transmission, but the texture of the screen is now visible.



a) b) c)

Figure 3.8: a) image of the test object, after mask centering. b) small test object, MURA 61 × 61, glass screen image c) small test object, MURA 61 × 61, paper screen image

3.3.3 Comparison to pinhole

For comparison purposes a pinhole image of the small test object was also taken. The image is not only much noisier than the coded aperture image but it also shows some artifacts because not all holes are equally bright, despite having the same size. Resolution is not as good as the code aperture image. This is because to generate some sort of image with the same source and exposure time used for the coded aperture, we needed to use a pinhole much wider than the mask pinhole size.

3.4 Validation with E-Cam data

All of the above simulations required the full availability of the imaging equipment and the possibility of building a number of masks in a short time. These trials could not have been done on the actual system. By courtesy of the Brigham and Women's hospital we were able to work on a Siemens E-Cam for a whole day. It was possible to compare the computer code directly with data from the E-Cam. In the past, our research group had built a 1.5-mm-thick lead 11 × 13 URA with mask pixels of 5 mm. The mask had a 1-mm-thick supporting aluminum backing plate. To reproduce results obtained by the

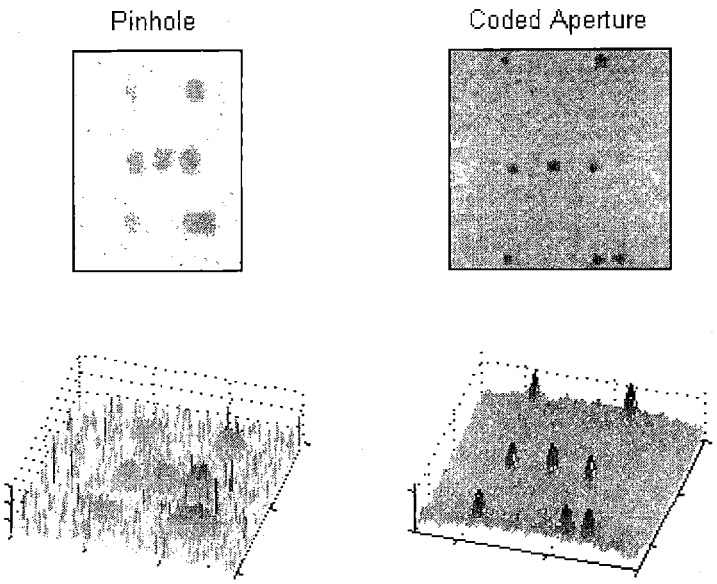


Figure 3.9: the object and the coded aperture image are the same of Figure 3.8. The pinhole image is affected by near field and random noise much more than the coded aperture. The exposure time was the same, but the pinhole had to be made larger than the pinholes of the coded aperture mask to get enough signal to take a picture.

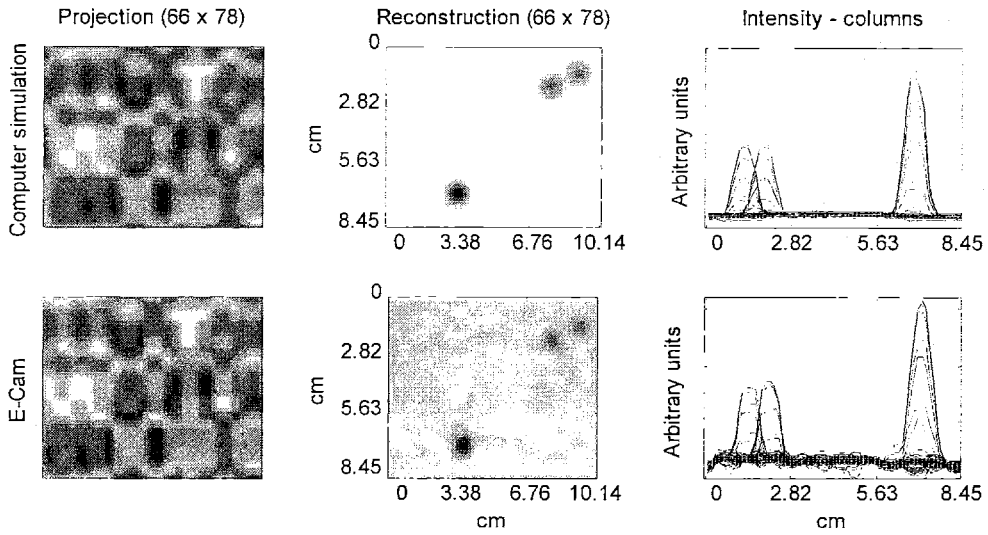


Figure 3.10: comparison of computer simulation with data from the E-Cam for a URA 11 × 13 pattern. One pixel is equivalent to 1.28 mm. The dimension of the points is 8.9 mm, against a geometric resolution of 7.7 mm. Object-to-detector distance: 28 cm. Near-field geometry effects appeared to be negligible. Relative intensities are reconstructed accurately. For the simulation α is integer and the mask is infinitely thin. Better agreement is possible if these parameters are changed.

designer of this mask, a mask-to-object distance of 10.1 cm and an object-to-detector distance of 19.1 cm were used. This implies $m = 2.89$, which leads (eq. (2.87)) to a resolution of 7.6 mm, a used detector area of 15.9 × 18.8 cm, and a field of view of about 8.5 × 10.1 cm. The maximum achievable resolution for this mask and the same field of view is obtained for maximum magnification, which is, from eq. (2.79):

$$m = \frac{d_d}{FoV} + 1 = \frac{38.7}{8.5} + 1 = 5.55 \tag{3.13}$$

because 53.3 × 38.7 cm is the maximum active area of the E-Cam[11]. Resolution would be, from eq. (2.87), 6.0 mm. Note that this is not the maximum resolution configuration for which we would have:

$$m = \frac{d_d}{d_m} = \frac{38.7}{0.5 \cdot 11} = 7.04 \tag{3.14}$$

which leads to 5.8 mm and a smaller field of view (7.58 × 6.4 cm). The aperture was tested with an object made of three point sources. Agreement with the computer simulation is remarkable (Figure 3.10).

[11] For the other dimension: $m = \frac{d_d}{h} + 1 = \frac{53.3}{10} + 1 = 6.33$, which is less restrictive.

Chapter 4 THE SIGNAL-TO-NOISE RATIO IN CODED APERTURE IMAGING

The main motivation behind coded aperture imaging is to improve the SNR relative to a pinhole camera.

The SNR properties of the URAs were calculated shortly after their application to imaging was proposed ([52]). These fundamental results showed that coded apertures do not always perform better than pinholes and that the biggest advantage is obtained for high background situations and point-like objects. Also, for extended objects, the SNR was showed to depend on the open fraction of the coded aperture, low fractions giving higher SNR. This led some researchers to suggest that low-throughput arrays (in particular product arrays) were suitable for use in Nuclear Medicine problems ([39]), where objects are typically extended.

To verify this statement we simulated the performance of a few families of masks. Results were surprising: not only did low-throughput arrays not outperform URAs, but they were also outperformed by other designs. This Chapter starts by showing these results. They triggered a reinvestigation of the SNR formulae available in literature: in particular an extension of the results of ref. ([52]) to arrays other than URAs was needed. After a precise definition of SNR is given, an expression valid for any family is derived. Particular care is dedicated to the issues of rescaling the decoding coefficients and mask transmission, neither of which is treated in the literature.

In section 4.3 the formula is specialized to different mask families. We found that our results, apart from minor corrections, confirm some particular cases available in literature ([37], [52]), to which they are compared. When applied to low-throughput arrays, the formula predicts the poor performance of product arrays observed in the simulations.

The Chapter ends with a discussion of problems in which coded apertures can be applied with considerable advantage over conventional techniques. Its main result is the identification of the array family chosen for the design of the prototype mask.

4.1 An intuitive summary of the problem

The idea behind code aperture imaging is making more efficient use of the detector. In fact, a detector with, say, $128 \times 128 = 16384$ pixels, can be thought as a set of 16384 independent counters. If the object were a point source, only one of these would be used in a pinhole imager. Coded apertures aim at getting better statistics by using more detectors, one detector for each open position, to take several measurements of the same source.

The potential SNR advantage over a pinhole is $\sqrt{N}$, with N number of open holes in the coded aperture. In fact, if the source is measured with a pinhole for a certain time giving the number of counts s (the signal), according to Poisson statistics, the associated standard deviation (the noise) is $\sqrt{s}$ and the SNR $\sqrt{s}$. If the same source is counted independently N times, the total count is, on the average, $N\,s$; the variance, being the sum of the variances of independent counts, is also $N\,s$, so that the SNR is $\sqrt{N\,s}$. The SNR advantage, the ratio of the SNR obtained with a coded aperture to the SNR obtained with a reference method, is, then, $\sqrt{N}$. A similar analysis was first done in the late 1950s by Fellgett in the field of spectroscopy, hence, in this field, the SNR advantage is sometimes called the Fellgett advantage ([26]). In imaging the name multiplexing advantage is also common ([31]). The prototype mask we eventually built has 480 holes. This means an SNR advantage of about 22. However, from the argument above, the SNR of a pinhole camera is recognized to be proportional to the square root of time or activity (because s is a total number of counts), so, a better way of evaluating this advantage is to think that 480 holes are equivalent to a reduction in time (or activity) of 480.

Unfortunately, this is a limit achieved only by point sources. It is interesting to note that if we knew a priori that the object is point source, and the objective is to determine its total brightness, the best strategy would be to use no mask at all, and simply take a count with the maximum possible open area. In our case, this would mean to open all the 62×62 positions of the mask, for a total of 3844 holes. Of course, it would not be fair to compare this result with an image because this integral count would not give the position of the point source. This shortcoming can be overcome using a single opaque element. The position of its shadow on the detector would sacrifice one pixel only, but give complete spatial information. This imager would be, so to speak, an anti-pinhole camera.

If the object is extended, the advantage is not as large. Several point sources project overlapping mask patterns, so each detector records information from more than one source and measurements are not independent. Overlaps, i.e. multiplexing, cause some costs in terms of SNR, depending on the extent of superposition. We shall see that if an object takes the whole field of view, the SNR advantage may be

lower than 1, i.e. a pinhole image would offer a better SNR, a result well known in literature ([31], [52]). Two conflicting needs arise: opening more pinholes, to increase the SNR advantage $\sqrt{N}$, and opening fewer pinholes, to reduce superposition. This trade-off is the necessary (but, as we shall see, not sufficient) condition for the existence of an optimal mask open fraction.

In conclusion, coded apertures try to take advantage of parts of the detector not used with conventional methods, provided that that such regions actually exist. This depends on the object and the open fraction of the aperture used. A qualitative analysis is needed to find if an optimal value of the open fraction actually exist. In the following, this analysis is presented. Other considerations, such as mask transmission, decoding coefficients, effect of background and different mask families are also taken into account.

4.2 SNR definition and choice of the decoding coefficients

The ultimate figure of merit of a medical image is the detectability of some abnormality (reference here). However, such measurements are very impractical and it is very common practice to rely on some definition of the Signal-to-Noise Ratio to have a quantitative measurement of image quality. This approach does have the limitation of depending on the particular definition of the SNR assumed. However, it is the method followed by the vast majority of the related literature. We never relied solely on the analytical results obtained, but constantly checked them against a visual assessment of simulated reconstructions, because we were aware of the problem

The SNR can be defined in a number of ways. After several trials, mathematical tractability and agreement with visual evidence were obtained for the simplest definition of SNR. Unlike detectability, our definition the SNR, which is essentially equivalent to that given in most literature (e.g. [3], [4], [37] and [52]), is not an overall figure of merit, but is defined on a local basis and varies point by point. At a given point, the SNR is defined to be the average value of the reconstructed activity at that pixel, divided by the standard deviation of this value. This definition, however, requires a little refinement. In fact, in these terms, the SNR would change if the image underwent a linear transformation. Let c be a number of counts collected in a Poisson process with mean n. According to the definition, "signal" is $n = c$, and "noise" its standard deviation $\sqrt{n}$, so that the SNR is $\sqrt{n}$. If this same calculation is carried out after a linear transformation of c, the "signal" becomes $s\,n + q$, and "noise" $s\sqrt{n}$, so that:

$$SNR = \frac{s \cdot n + q}{\sqrt{s^2 n}} = \sqrt{n} + \frac{q}{s\sqrt{n}} \tag{4.1}$$

with s and q constants. The SNR depends on s and q, which is not acceptable.

In section 2.7 was shown that the decoding array can be scaled in a number of ways to accommodate different needs. The scaling of $\mathbf{G}$ does not influence image reconstruction. In fact, if $\mathbf{G}^{\delta}$ is the decoding array such that $\mathbf{A} \otimes \mathbf{G}^0 = N\delta$ (where $\mathbf{A}$, having a physical meaning, is an array of 1s and 0s and δ is a true discrete δ function, i.e. is 1 on the peak and 0 outside), decoding with the array $\mathbf{G}^{lin} = s\mathbf{G}^0 + q\mathbf{1}$, where $\mathbf{1}$ is an array of 1s of the same size as $\mathbf{G}^0$, gives:

$$\mathbf{A} \otimes \mathbf{G}^{lin} = \sum \mathbf{A}\left(s\mathbf{G}^0 + \mathbf{1}q\right) = s\sum \mathbf{A}\mathbf{G}^0 + q\sum \mathbf{A}\mathbf{1} = sN\,\delta + qN \tag{4.2}$$

which is a δ-like function. Choosing s and q is equivalent to choosing the g_+ and g_- of section 2.7. To see the impact on the image, consider eq. (1.5). If $\mathbf{G}^0$ is replaced by any of its linear transformations $\mathbf{G}^{lin}$, from the linearity of the operators, the reconstructed image, $\hat{\mathbf{O}}^{lin}$, is just a linear transformation of $\hat{\mathbf{O}}$, i.e. $\hat{\mathbf{O}}^{lin} = sN\hat{\mathbf{O}} + qN$. This is a change in the brightness and contrast of the image, typically overridden by the displaying equipment which rescales the data, leading to the same result obtained with $\mathbf{G}^0$. Provided that the correct correlation is used, there is no difference in looking at an image after decoding with $\mathbf{G}^0$ or any of the $\mathbf{G}^{lin}$s. From a theoretical point of view, this is important because if $\mathbf{G}^0$ is a valid decoding array, so is $\mathbf{G}^{lin}$, which makes the constants s and q available for some kind of optimization, examples of which were given in section 2.7. One of the objectives of the calculation of the SNR of a coded aperture is to find, in the same fashion, the pair (g_+, g_-) that avoids the appearance of spurious terms, like the term proportional to q in eq. (4.1), in the SNR.

The calculation is simple but involved and must be carried out under a number of approximations. Far-field geometry is assumed. Furthermore, each detector pixel has the same size of the projection of a mask hole on the detector, which is assumed to fall exactly on detector a pixel. To keep track of all constants, the decoding process is followed with the decoding array $\mathbf{G}^{lin}$ rather than with $\mathbf{G}^0$. Some new variables are introduced to handle the effects of background, statistics and partial mask transmission. Trying to keep the notation consistent with that of ref. [52], let (u,v) be a position of the detector and:

Ω_{kl}: matrix of the number of the photons collected at (u,v), due to the sources present within the kl^{th} reconstruction position and emitted in a solid angle subtending a single mask pixel. In a far-field approximation, the elements of the kl^{th} matrix all are different

only because they are different realizations of the same random process (the decay of the kl^{th} source) and have the same average value.

$'\Omega_{kl}$: matrix of the number of photons collected at (u,v), due to the sources present at the kl^{th} reconstruction position, emitted in an angle subtending a single mask pixel, after passing through a hypothetical mask with all positions closed (but partially transparent).

O: matrix of the average activity of the object: its entry (k,l) is defined as the mean value of any of the elements of Ω_{kl}. It is a function of (u,v) only.

I_T sum of all the elements of **O**. It is proportional to the total power of the object.

D: matrix of counts due to uncoded noise at (u,v).

B: mean value of **D**. Since **D** is assumed uniform over the detector, B is a scalar quantity.

N_T total number of elements in the mask.

N total number of open elements in the mask.

t: mask transmission (average fraction of photons that passes through an opaque element without interaction)

4.2.1 Definition of signal

With these definitions, **R**, the number of counts recorded at (u,v) is:

$$\mathbf{R}(u,v) = \sum_{kl} \mathbf{\Omega}_{kl}(u,v)\mathbf{A}(u+k,v+l) + \sum_{kl} {}'\mathbf{\Omega}_{kl}(u,v)\left[\mathbf{1} - \mathbf{A}(u+k,v+l)\right] + \mathbf{D}(u,v) \tag{4.3}$$

Since all functions involved are functions of (u,v), the argument is henceforth omitted. The dependence of **A** and **G** on shift is indicated with a subscript: $\mathbf{A}_{kl}$ indicates the pattern **A**, shifted to represent the projection of a point source located at reconstruction position (k,l) and $\mathbf{G}_{ij}$ indicates the array **G** shifted to decode position (i,j) of the reconstructed image.

In eq. (4.3), **R** is a particular realization of a random process because, Ω_{kl}, $'\Omega_{kl}$ and **D** are Poisson-distributed random variables. This is obvious for **D**, which is background, and Ω_{kl}, which is due to γ-rays coming directly from the different sources in the object. However, $'\Omega_{kl}$ represents γ-rays reaching the detector after passing through shielding and is the cascade of a Poisson process (the radioactive emission

associated with Ω_{kl}) with a binomial process (passing with probability t through the mask): ${}^{t}\Omega_{kl}$ is, then, Poisson distributed (see Appendix C.1), with mean $t\,\mathbf{O}$. $\mathbf{R}$ is the sum of Poisson variables and is, thus, Poisson distributed.

The reconstructed image is given by eq. (1.5):

$$\hat{\mathbf{O}}(i,j) = \mathbf{R} \otimes \mathbf{G}_{ij}^{lin} = \sum_{u,v}\sum_{k,l} \mathbf{\Omega}_{kl}\mathbf{A}_{kl}\left(s\mathbf{G}_{ij}+q\right) + \sum_{u,v}\sum_{k,l} {}^{t}\mathbf{\Omega}_{kl}\left(\mathbf{1}-\mathbf{A}_{kl}\right)\left(s\mathbf{G}_{ij}+q\right) + \sum_{u,v}\mathbf{D}\left(s\mathbf{G}_{ij}+q\right) \quad (4.4)$$

Rearranging:

$$\begin{aligned}\hat{\mathbf{O}}(i,j) = s\left[\sum_{u,v}\sum_{k,l}\left(\mathbf{\Omega}_{kl}\mathbf{A}_{kl}\mathbf{G}_{ij} - {}^{t}\mathbf{\Omega}_{kl}\mathbf{A}_{kl}\mathbf{G}_{ij} + {}^{t}\mathbf{\Omega}_{kl}\mathbf{1}\mathbf{G}_{ij}\right) + \sum_{u,v}\mathbf{D}\mathbf{G}_{ij}\right] + \\ + q\left[\sum_{u,v}\sum_{k,l}\left(\mathbf{\Omega}_{kl}\mathbf{A}_{kl} - {}^{t}\mathbf{\Omega}_{kl}\mathbf{A}_{kl} + {}^{t}\mathbf{\Omega}_{kl}\mathbf{1}\right) + \sum_{u,v}\mathbf{D}\right]\end{aligned} \quad (4.5)$$

In these sums all terms have random variables contributing variance to $\hat{\mathbf{O}}(i,j)$. The terms multiplied by s all depend on (i,j) and, thus, vary across the image. The expectation value of this part is:

$$\begin{aligned}E\left[\hat{\mathbf{O}}^{\mathrm{mod}}(i,j)\right] &= s\left[\sum_{u,v}\sum_{k,l}\left(\mathbf{O}(k,l)\mathbf{A}_{kl}\mathbf{G}_{ij} - t\mathbf{O}(k,l)\mathbf{A}_{kl}\mathbf{G}_{ij} + t\mathbf{O}(k,l)\mathbf{1}\mathbf{G}_{ij}\right) + \sum_{u,v} B\mathbf{G}_{ij}\right] \\ &= s\left[(1-t)\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij} + t\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{G}_{ij} + B\sum_{u,v}\mathbf{G}_{ij}\right]\end{aligned} \quad (4.6)$$

Now assume that $\mathbf{G}$ is $\mathbf{G}^0$, i.e. the version of the decoding array for which: $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^{0} = N\delta(i\text{-}k,j\text{-}l)$. With this, the equation above becomes:

$$E\left[\hat{\mathbf{O}}^{\mathrm{mod}}(i,j)\right] = s(1-t)N\mathbf{O}(i,j) + s(tI_T + B)\sum_{u,v}\mathbf{G}_{ij}^{0} \quad (4.7)$$

$\sum_{u,v}\mathbf{G}_{ij}^{0}$ does not depend on the reconstruction position (i,j). The first term represents the net signal. The effect of mask transmission is to decrease the signal by the amount of counts that have passed through the mask. This is quite reasonable, because all these counts partially fill positions that would be dark in an ideal case, giving the same number of counts as a reduced source over an increased constant

background. And in fact, in the second term, transmission background (tI_T) and uncoded background (B) sum to the same flat (in an average sense only) contribution.

In eq. (4.5), the terms multiplied by q do not depend on (i,j) and must be exactly constant (not only in an average sense) over the whole image. They add a *perfectly* flat background, which does fluctuate over different images, but does not change in the same image because it affects all pixels in the same way. Accordingly, even if they contribute variance to the absolute value of $\hat{\mathbf{O}}(i,j)$, they should not be included in the noise term either. The origin of this flat pedestal can be traced back by taking the expectation value:

$$\begin{aligned} E\left[\hat{\mathbf{O}}^{ped}(i,j)\right] &= q\left[\sum_{u,v}\sum_{k,l}\left(\mathbf{O}(k,l)\mathbf{A}_{kl} - t\mathbf{O}(k,l)\mathbf{A}_{kl} + t\mathbf{O}(k,l)\mathbf{1}\right) + \sum_{u,v} B\right] = \\ &= q\left[(1-t)\sum_{u,v}\sum_{k,l}\mathbf{O}(k,l)\mathbf{A}_{kl} + t\sum_{u,v}\sum_{k,l}\mathbf{O}(k,l) + N_T B\right] = \\ &= q\left[(1-t)\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl} + t\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{1} + N_T B\right] = \\ &= q\left[(1-t)NI_T + tN_T I_T + N_T B\right] \end{aligned} \tag{4.8}$$

The first term comes from the net signal (it is the contribution of the sidelobes), the second from mask transmission, the third from unencoded background counts. $\mathbf{G}^0$ is a convenient choice of $\mathbf{G}$ because $q = 0$ and $E[\,\hat{\mathbf{O}}^{ped}(i,j)\,] = 0$, i.e. the pedestal and its fake variance do not appear in $\hat{\mathbf{O}}(i,j)$.

In conclusion, the decoding of the collected counts $\mathbf{R}$ with an arbitrarily scaled decoding array $\mathbf{G}^{lin}$ is the sum of three kinds of terms: one proportional to the emitting source and two flat contributions. Of these, one is flat in an average sense only while the other is perfectly constant. The suppression of these last two terms would require, respectively, $s = 0$ and $q = 0$ (unless $\sum_{u,v}\mathbf{G}^0_{ij} = 0$, which would make any s acceptable, provided $q = 0$, but may not be the case), which is clearly unacceptable. The first two contributions add noise to the image and share the property of being proportional to s. The third provides a pedestal level, proportional to q, which fluctuates over different images but is uniform on the same image, so that it introduces no visible noise. In light of this, the signal is defined to be:

$$E[\hat{\mathbf{O}}] \equiv (1\text{-}t)\, s\, N\, \mathbf{O} \tag{4.9}$$

4.2.2 Definition of noise

Given the definition of signal of the previous section, it makes sense to identify noise with the standard deviation of the term proportional to s in eq. (4.5). The calculation starts again from the reconstructed image of eq. (4.4) and calculate the variance. This expression is a convenient starting point because all random variables in here (Ω_{kl}, $'\Omega_{kl}$ and $\mathbf{D}$) are independent and appear only once. The variance is then the variance of the sum of independent variables, which is the sum of the variances. Using the fact that $\mathbf{A}$ is binary with values 0 and 1, and thus $\mathbf{A} = \mathbf{A}^2$:

$$\mathrm{var}\left[\hat{\mathbf{O}}(i,j)\right] = \sum_{u,v}\sum_{k,l}\mathbf{O}(k,l)\mathbf{A}_{kl}\left(s\mathbf{G}^0_{ij}+q\right)^2 + \sum_{u,v}\sum_{k,l}t\mathbf{O}(k,l)\left(\mathbf{1}-\mathbf{A}_{kl}\right)\left(s\mathbf{G}^0_{ij}+q\right)^2 + \sum_{u,v}B\left(s\mathbf{G}^0_{ij}+q\right)^2 \quad (4.10)$$

Unlike Ω_{kl},, $\mathbf{O}(k,l)$ is a constant over the detector and can be taken outside the summation over u and v. A similar observation applies to $\mathbf{D}$, leading to:

$$\mathrm{var}\left[\hat{\mathbf{O}}(i,j)\right] = (1-t)\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\left(s\mathbf{G}^0_{ij}+q\right)^2 + t\sum_{k,l}\mathbf{O}(k,l)\mathbf{1}\sum_{u,v}\left(s\mathbf{G}^0_{ij}+q\right)^2 + B\sum_{u,v}\left(s\mathbf{G}^0_{ij}+q\right)^2 \quad (4.11)$$

Simple algebra gives:

$$\begin{aligned}\mathrm{var}\left[\hat{\mathbf{O}}(i,j)\right] = {} & (1-t)s^2\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}^{0\,2}_{ij} + 2(1-t)sq\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}^0_{ij} + (1-t)q^2\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl} + \\ & s^2t\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{G}^{0\,2}_{ij} + 2sqt\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{G}^0_{ij} + t\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{1}q^2 + \\ & s^2B\sum_{u,v}\mathbf{G}^{0\,2}_{ij} + 2sqB\sum_{u,v}\mathbf{G}^0_{ij} + q^2\sum_{u,v}B\end{aligned} \quad (4.12)$$

Using $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}^0_{ij} = N\delta(i-k,j-l)$, $\sum_{u,v}\mathbf{A}_{kl} = N$, $\sum_{k,l}\mathbf{O}(k,l) = I_T$, $\sum_{u,v}q^2\mathbf{1} = q^2N_T$, $\sum_{u,v}B = N_TB$

and observing that $\sum_{u,v}\mathbf{G}^0_{ij}$ is a constant, not dependent on (k,l) or the shift (i,l):

$$\begin{aligned}\mathrm{var}\left[\hat{\mathbf{O}}(i,j)\right] = {} & (1-t)s^2\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}^{0\,2}_{ij} + 2(1-t)sqN\mathbf{O}(i,j) + (1-t)q^2NI_T + \\ & s^2t\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{G}^{0\,2}_{ij} + 2sqtI_T\sum_{u,v}\mathbf{G}^0_{ij} + tN_Tq^2I_T + \\ & s^2B\sum_{u,v}\mathbf{G}^{0\,2}_{ij} + 2sqB\sum_{u,v}\mathbf{G}^0_{ij} + q^2N_TB\end{aligned} \quad (4.13)$$

One can also start from eq. (4.5), which is just eq. (4.4) rearranged, but has to be careful because its two terms are not statistically independent because the same realizations of the random variables appear in different places. The result is the same as eq. (4.13), and has the form:

$$\mathrm{var}\left\{\hat{\mathbf{O}}(i,j)\right\} = \mathrm{var}\{s[...]\} + 2\,\mathrm{cov}\{s[...], q[...]\} + \mathrm{var}\{q[...]\} \tag{4.14}$$

where in the square brackets are the same terms that appear in eq. (4.5). With a little work one can confirm that this is the same as eq. (4.13), but this is not necessary to prove that the first column of eq. (4.13) is the variance of the term proportional to s in eq. (4.5), the second is the covariance term and the third is the variance of the term proportional to q, because one can just look at the constants s and q to reach the same conclusion. After the discussion on the origin of the variance of the signal, we know that the only variance of interest is the variance of the term proportional to s because other terms are the variance of a perfectly flat contribution that varies only over different images, and consequently define noise as:

$$\mathbf{N} = s\sqrt{(1-t)\sum_{k,l}\mathbf{O}(i,j)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^{0\,2} + (tI_T + B)\sum_{u,v}\mathbf{G}_{ij}^{0\,2}} \tag{4.15}$$

where the contributions are, from left to right, from statistical noise in the net signal, transmission and uncoded background. Also note that s, also present in the signal, cancels out in the SNR:

$$SNR(i,j) = \frac{(1-t)N\,\mathbf{O}(i,j)}{\sqrt{(1-t)\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^{0\,2} + (tI_T + B)\sum_{u,v}\mathbf{G}_{ij}^{0\,2}}} \tag{4.16}$$

This means that s can be arbitrarily chosen. With the current definitions, choosing $s = 1$ implies that a unit point source reconstructs to a peak of height N, which is quite convenient, because it is consistent with the definition that a unit point source reconstructs to a peak of height 1 in a pinhole imager. From the definition of signal (eq (4.9)), this choice is the one that conserves the number of counts in the image in the ideal case of no mask transmission, i.e. $\sum_{i,j}\hat{\mathbf{O}}(i,j) = \sum_{u,v}\mathbf{R}(u,v)$. The final step is to divide numerator and denominator by the total source power to introduce two dimensionless parameters, $\psi_{ij} = \frac{\mathbf{O}(i,j)}{I_T}$ and $\xi = \frac{B}{I_T}$. ψ and ξ express, respectively, the fraction of the object activity concentrated at one reconstruction position (in this sense ψ is a concentration parameter) and the background level (at

one pixel) relative to the total object activity. ψ is constrained between 0 and 1 while ξ is always positive. Two interesting limiting cases are that of a point source and an uniform object. In the first case, all activity is concentrated at one point, so $\psi = 1$ at one point and 0 elsewhere; in the second, it is uniformly spread over N_T, so $\psi = 1 / N_T$ everywhere. Note that for points with no activity $\psi = 0$ and SNR = 0. With the new definitions:

$$\begin{cases} SNR(i,j) = \dfrac{\sqrt{I_T}(1-t)N\psi_{ij}}{\sqrt{(1-t)\sum_{k,l}\psi_{kl}\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2 + (t+\xi)\sum_{u,v}\mathbf{G}_{ij}^2}} \\ \mathbf{G} = \mathbf{G}^0 \end{cases} \tag{4.17}$$

From eq. (4.17) we can see that, once a pattern and its decoding array are given, the quantities $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2$ and $\sum_{u,v}\mathbf{G}_{ij}^2$, which is sometimes called the energy of the decoding pattern ([53]), are all that is needed to calculate the SNR. From the derivation it is clear that the formula holds only if $\mathbf{G} = \mathbf{G}^0$, i.e. is such that $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij} \propto N\delta$. This condition is equivalent to choosing some particular linear transformations of $\mathbf{G}$, those for which $q = 0$, and is emphasized in eq. (4.17) to be part of the definition of the SNR. In fact, linear transformations of $\mathbf{G}$ such that $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij} = sN\delta + q$ introduce terms that should be disregarded in the calculation of the SNR: eq. (4.7), (4.8), and (4.17) provide the terms that would have to be subtracted from an image reconstructed with such $\mathbf{G}$s and its variance before the SNR is calculated as the ratio.

The SNR changes from point to point. The only literature example we know of a global definition is ref. [53], where noise and signal are summed over the whole image. This was expressly done to simplify the expression of the SNR by eliminating dependence on the parameter ψ, a step which, as we shall see, would not be to our advantage.

Except for the considerations on linear transformations and mask transparency, a similar formula was reached by Gottesman and Schneid with a heuristic derivation ([37]). In the next section it is specialized to some families and is shown to lead to very intuitive results for some simple cases, such as the pinhole camera or a point source.

4.3 The Signal-to-Noise Ratio of different coded apertures

In this section is calculated the SNR of some coded aperture families. The most important is the case of the (M)URA, because it provides a reference point for comparison with literature results. The case of product arrays is then treated to show that this family always performs worse than (M)URAs. The mask we eventually built is a NTHT array, which is treated next. The case of the pinhole is also indispensable for the calculation of the SNR advantage. The interesting case of the negative pinhole closes this section providing additional insight.

From the definition of the SNR, the first step is always to find the particular $\mathbf{G} = \mathbf{G}^0$.

4.3.1 Uniformly and Modified Uniformly Redundant Arrays

For MURAs and URAs the number of open mask positions is, respectively, $N = (N_T + 1) / 2$ and $N = (N_T - 1) / 2$. In matched decoding the sidelobes have a value M, which is useful in our calculations because it represents the number of 1s in **A** that overlap with g_+'s in **G** for shifts (relative to **A**) other than 0. Applying these properties:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij} = Mg_+ + (N-M)g_- + \delta(i\text{-}k,j\text{-}l)(N-M)(g_+ - g_-) \tag{4.18}$$

Setting $\mathbf{G} = \mathbf{G}^0$ gives $g_+ = 1$ and $g_- = M / (M - N)$, which are the coefficients of balanced decoding. For a 50% open array $M = N / 2$ so that $g_- = -1$. For the more general case of an arbitrary open fraction ρ, $N = \rho N_T$. In the approximation $M \cong \rho^2 N_T$, $g_- = \rho / (\rho - 1)$:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij}^2 = \frac{\rho N_T}{1-\rho}\left[\rho + \delta(i-k,j-l)(1-2\rho)\right] \quad \text{and} \quad \sum_{u,v} \mathbf{G}_{ij}^2 = \frac{\rho N_T}{1-\rho} \tag{4.19}$$

which substituted in eq. (4.17), gives the SNR:

$$SNR_{ij} = \frac{\sqrt{N_T I_T}\sqrt{\rho(1-\rho)}(1-t)\psi_{ij}}{\sqrt{(1-t)\left[\rho + (1-2\rho)\psi_{ij}\right] + t + \xi}} \tag{4.20}$$

which is the same as Fenimore's expression ([52]), except for the factor 2 in the denominator. As all definitions are consistent, this difference comes from the substitution, in his derivation, of a sum over all sources except $\mathbf{O}(i,j)$ with I_T, causing an overestimate of the variance not relevant for most cases but with some effects for large ψ. An example is the case of a point source. With no transmission or uncoded

background ($\psi_{ij} = \delta(i,j)$, $t = 0$ and $\xi = 0$), the SNR at the source location is $\sqrt{\rho N_T I_T} = \sqrt{N I_T}$, which makes sense, because under these conditions all the coded aperture does is to count the same source N times. There is no need to justify an apparent $\sqrt{2}$ reduction in SNR as many authors do.

The SNR is always positive and is zero for a completely open and a completely closed pattern. Setting the first derivative of the SNR with respect to ρ to 0 gives the open fraction that maximizes the SNR:

$$\rho_{opt,ij} = \frac{t + \xi + \psi_{ij}(1-2t) - \sqrt{[t + \xi + \psi_{ij}(1-2t)]^2 - (1-t)(2\psi_{ij} - 1)[\psi_{ij}(1-t) + t + \xi]}}{(1-t)(2\psi_{ij} - 1)} \tag{4.21}$$

which, for $t = 0$, reduces to:

$$\rho_{ij}^{opt} = \frac{\xi + \psi_{ij} - \sqrt{(\xi + \psi_{ij})(1 + \xi - \psi_{ij})}}{2\psi_{ij} - 1} \tag{4.22}$$

ρ_{ij}^{opt} can range from 0 to 1 (Figure 4.1), depending on the activity concentration at the point of interest in the image. If it existed, the best (M)URA to image a point-like source ($\psi_{ij} = 1$) with no background ($\xi = 0$) would be a completely open pattern, which, as already discussed, is not surprising. Note that the existence of ρ^{opt} does not have anything to do with the actual existence of a corresponding

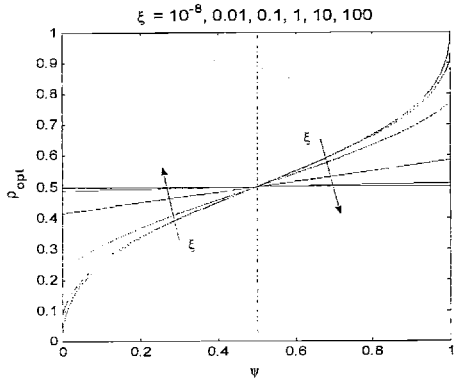


Figure 4.1: optimal open fraction as a function of ψ for different backgrounds for (M)URAs. (eq. 4.22). Note that especially at low background, the optimal open fraction for low ψ_{ij} (extended objects) can be significantly less than 0.5. It can be proved analytically that all curves have odd symmetry about (0.5, 0.5)

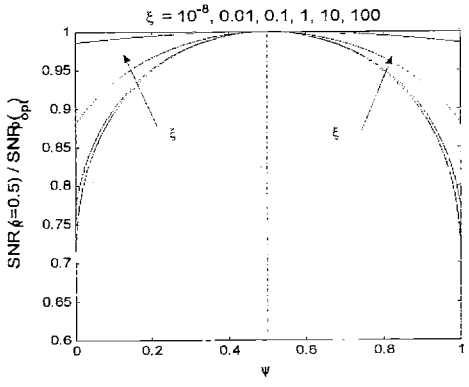


Figure 4.2: SNR loss in using the half-open pattern in place of the optimal, for different backgrounds. The loss is never more than 25%, as first reported ([52]). Following from the error pointed out in the text, in this reference these graphs stopped at $\psi_{ij} = 0.5$, about which curves are symmetric.

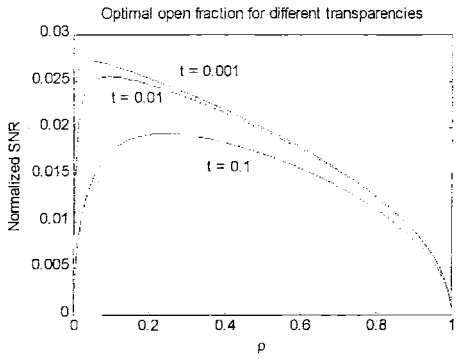


Figure 4.3: SNR as a function of the open fraction for different mask transmission (0.1, 0.01 and 0.001). The result is normalized to the SNR of the pinhole and $\sqrt{N_T}$. For the case of our interest, t = 0.01. Since the maximum normalized SNR is 0.025, N_T must be at least 40 for the coded aperture to perform better than the pinhole.

(M)URA. In fact a totally open (M)URA does not exist because a totally open mask can convey no spatial information. This result is not discussed in ref. [52] (or in the anywhere in literature) because, following the algebraic mistake pointed out above, graphs ended up stopping at $\psi_{ij} = 0.5$.

In images where more than half of the activity is concentrated in one reconstruction pixel, more than half open (M)URAs should be preferred if that one pixel is the only one of interest. To our knowledge, these patterns do not exist in two dimensions, but more than half-open 1d URA sequences can be constructed by inverting dilute URAs (§2.4.3). It must be pointed out that best performance is obtained at the point for which $\psi_{ij} > 0.5$, and that, since $\Sigma_{ij} \psi_{ij} = 1$, at most one pixel can have $\psi_{ij} > 0.5$. Also, the greater the maximum ψ_{ij}, the lower all other ψ_{ij}'s and the greater the loss in SNR for all other points. In the vast majority of cases high-open fraction apertures are not beneficial.

A second interesting implication of eq. (4.22) is that the higher the background, the lower the influence of choosing the right open fraction, as it can be verified in Figure 4.1. In Nuclear Medicine applications this is not the case: typical backgrounds are as low as $\xi \sim 5 \times 10^{-4}$. Since typical average ψ_{ij}'s are about 0.01, the loss, though at worst 25% in SNR, still amounts to a factor of 2 in exposure times or activity. Hence our efforts to find an optimal pattern.

With the parameters applicable to the thyroid phantom ($\psi_{ij} = 2.6 \times 10^{-4}$, $\xi = 5.5 \times 10^{-4}$), assuming a transmission of 1% (see §6.1.4) one can find that the optimum open fraction is 9.42%. In Figure 4.3 this value is shown to depend on mask transmission and increases with increasing transmission. Because of this result, we set out to find a 10% open pattern.

4.3.2 Product arrays

For product arrays (§2.4.5) $\mathbf{A} \neq \mathbf{G}$, which was an assumption necessary to derive the results of §4.3.1. This is the reason they can not be applied to product arrays and one must restart from eq. (4.17). The definitions given in §2.4.5 lead directly to $\mathbf{G}^0$. The elements of this array are 1 and –1, which is enough to prove that:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij}^2 = N \quad \text{and} \quad \sum_{u,v} \mathbf{G}_{ij}^2 = N_T \tag{4.23}$$

Substitution in eq. (4.17) gives:

$$SNR_{ij} = \frac{\sqrt{N_T I_T}\rho(1-t)\psi_{ij}}{\sqrt{(1-t)\rho + t + \xi}} \tag{4.24}$$

Since for all physically allowable values of the parameters the SNR is an increasing function of ρ, the optimal open fraction is simply the largest that can be achieved. Substituting ρ = 0.25, which is the only possible value for PNP arrays as defined in ref. [37], the results of this paper are reproduced.

4.3.3 No-Two-Holes-Touching (M)URAs

With the rules of ref. [13] and [14] only half-open (M)URAs can be generated. This limitation can be overcome by inserting a number *e*-1 of opaque columns (rows) between all columns (rows) of an original array, to form No-Two-Holes-Touching (M)URAs (§2.4.7). The open fraction can only be smaller than that of the originating array and the mask pattern looks like a grid in which are placed the elements of the original array. The decoding array $\mathbf{G}^0$ is a three-valued function. At the positions of the original elements are present 1s and –1s, while at all other positions are present 0s. This structure does influence the quantities we have to calculate, which take on different values depending on whether or not the shift is such that the blank lines of **G** overlap with the non-blank lines of **A**:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij}^2 = \begin{cases} 0 & \text{for } N_T^0\left(e^2-1\right)\text{shifts} \\ N & \text{for } N_T^0 \text{ shifts} \end{cases} \quad \text{and} \quad \sum_{u,v} \mathbf{G}_{ij}^2 = N_T^0 \tag{4.25}$$

where N_T^0 is the total number of positions of the original array while $N_T = e^2\ N_T^0$ still is the total number of positions in the pattern. With this:

$$\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2 = N\sum_{k,l}^{N_T^0}\mathbf{O}(k,l) \tag{4.26}$$

where the summation on the right hand side indicates a sum over pixels of the image reconstructed for shifts of **G** which superimpose its blank lines to those of **A**. If the source has no particular structure (such as a point (see §7.4.2) or a line parallel to the lines of **G**), $\sum_{k,l}^{N_T^0}\mathbf{O}(k,l)$ is simply the normalized activity of a $1/e^2$ fraction of it. One can substitute this sum over a partial number of elements with a sum over all elements:

$$\sum_{k,l}\mathbf{O}(k,l)\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2 = \frac{N}{e^2}\sum_{k,l}^{N_T}\mathbf{O}(k,l) = \frac{N}{e^2}I_T \tag{4.27}$$

Since the only (M)URAs we know are 50% open, the calculation is carried on only in this case, for which the open fraction of the array becomes a function of e only: $\rho = (2\,e^2)^{-1}$. The final result is:

$$SNR_{ij} = \frac{\sqrt{N_T I_T}\sqrt{\rho/2}(1-t)\psi_{ij}}{\sqrt{(1-t)\rho + t + \xi}} \tag{4.28}$$

Taking the derivative with respect to ρ proves that best performance is again obtained for maximum ρ, *i.e.* $\rho = 0.5$, which would be the same as the original URA: a 50% URA is always better than a NTHT array based on it. The maximum, non-trivial, ρ is obtained for $e = 2$ and, thus, $\rho_{opt} = 0.125$. A second important result is that, for any given open fraction $\rho \leq 0.5$, the SNR of a NTHT is higher than that of product arrays.

Note that for NTHT patterns a valid pattern is obtained by inverting the elements of the original (M)URA pattern only, but not the opaque lines. The pattern is still self-supporting. For existing 50% open patterns, this does not change the open fraction, so that the SNR is the same of the original NTHT pattern. Even if this pattern is not the negative of the original pattern, it still is effective in eliminating near-field artifacts (see Chapter 5).

4.3.4 Pinhole

The pinhole itself is a limiting case of NRA because $\mathbf{A} \otimes \mathbf{A} = \delta$. The method used for all arrays can be applied to calculate the SNR with $\mathbf{G} = \mathbf{A}$. One can see right away that:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij}^2 = \delta(i-k, j-l) \quad \text{and} \quad \sum_{u,v} \mathbf{G}_{ij}^2 = 1 \tag{4.29}$$

which lead to:

$$SNR_{ij} = \frac{\sqrt{I_T}(1-t)\psi_{ij}}{\sqrt{(1-t)\psi_{ij} + t + \xi}} \tag{4.30}$$

in complete agreement with a direct calculation and ref. [37] and [52][12], providing an additional validation of eq. (4.17). Note that this formula may be unfair to the pinhole because it assumes uniform mask thickness; however a pinhole mask can be made very thick far from the pinhole, reducing transmission. For this reason, very often in calculations it is better to set $t = 0$ for the pinhole but not for coded apertures.

4.3.5 The negative pinhole

By "negative pinhole" we indicate a mask completely open, except for one location. The appropriate decoding array is obtained by setting $\mathbf{A} \otimes \mathbf{G}^{\delta} = (N_T\text{-}1)\delta$, which leads to $g_+ = 2\text{-}N_T$ and $g_- = 1$. With these:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij}^2 = (N_T - 1)\left[N_T - 2 + (3 - N_T)\delta(i-k, j-l)\right] \quad \text{and} \quad \sum_{u,v} \mathbf{G}_{ij}^2 = N_T^2 - 3N_T + 3 \tag{4.31}$$

which lead to, with $N = N_T - 1$:

$$SNR_{ij} = \frac{\sqrt{\dfrac{N_T - 1}{N_T - 2}}\, I_T (1-t)\psi_{ij}}{\sqrt{(1-t)\left(1 - \dfrac{N_T - 3}{N_T - 2}\psi_{ij}\right) + (t+\xi)\dfrac{N_T^2 - 3N_T + 3}{N_T^2 - 3N_T + 2}}} \tag{4.32}$$

The expression can be simplified with the approximation $N_T >> 1$:

$$SNR_{ij} = \frac{\sqrt{I_T}(1-t)\psi_{ij}}{\sqrt{1 - (1-t)\psi_{ij} + \xi}} \tag{4.33}$$

The SNR advantage over the pinhole is:

$$SNR_{NP/P} = \frac{\sqrt{(1-t)\psi_{ij} + t + \xi}}{\sqrt{1-(1-t)\psi_{ij} + \xi}} \tag{4.34}$$

This is always an increasing function of ψ_{ij}. The breakeven point is $\psi_{ij} = 0.5$, $\forall$ t, ξ. If $\psi_{ij} < 0.5$, the pinhole prevails, while the opposite is true for $\psi_{ij} > 0.5$. It is interesting to see that for $\psi_{ij} = 0$ the advantage is $\sqrt{\frac{t+\xi}{1+\xi}}$ while for $\psi_{ij} = 1$ it is $\sqrt{\frac{1+\xi}{t+\xi}}$. If $\xi \to \infty$ or $t = 1$, both these limits, that confine the whole curve because it is monotonically increasing, tend to 1. For normal values of the parameters, however, $\xi \cong 0$ and t is a few percent, which makes the negative pinhole very attractive for imaging concentrated sources.

Unfortunately, the approximation $N_T >> 1$ does not hold for the important case $\psi_{ij} = 1$. In this case, for $t = 0$, one can prove that the SNR advantage over the pinhole is:

$$SNR_{NP/P} = \sqrt{N_T - 1}\,\frac{\sqrt{1+\xi}}{\sqrt{1+\xi\left(1+\frac{1}{N_T-1}\right)}} \tag{4.35}$$

which is always greater than 1. Thus, for a point source, a better *local* SNR is obtained with the negative pinhole. This does not mean necessarily a better image, because the SNR at the sidelobes may be much worse. However, *at* the point of interest, a better result is obtained. This apparent contradiction originates from the local nature of the definition of SNR.

A last case of interest is that of no background and no transmission. The negative pinhole has an advantage only for points with $\psi_{ij} > 0.5$. If $\psi_{ij} = 1$ the advantage is $\sqrt{N_T - 1}$ as can be obtained from eq. (4.35). Note that the other factor in eq. (4.35) represents the loss of advantage caused by background.

4.3.6 Random array

For the random array, the height of the peak is Ng_+. Setting this to N gives immediately $g_+ = 1$. The sidelobe value is given by $(N\text{-}M)g_- + Mg_+$ where M is the number of superpositions of 1s of **A** with

---

[12] In this reference the SNR of the pinhole was calculated with an independent calculation, so it is not affected by the above mentioned algebraic error.

g_+'s of **G** and can be approximated with ρN. Setting the sidelobe to 0 gives $g_- = \rho / (\rho - 1)$, which are the coefficients of balanced decoding. With these:

$$\sum_{u,v} \mathbf{A}_{kl}\mathbf{G}_{ij}^2 = \begin{cases} N & \text{for no shift} \\ \rho N \cdot 1^2 + (N - \rho N)\left(\frac{\rho}{\rho - 1}\right)^2 & \text{otherwise} \end{cases} \quad \text{and} \quad \sum_{u,v} \mathbf{G}_{ij}^2 = N + (N_T - N)\left(\frac{\rho}{\rho - 1}\right)^2 = N\frac{1}{1-\rho} \quad (4.36)$$

which are the same as the (M)URA case. The same SNR expression applies. While known (M)URAs only have 50% open fraction, random arrays can be made of any open fraction, so that the maximum SNR predicted for (M)URAs can actually be obtained. However, this SNR has to be combined with the inherent SNR (see §2.4.1). If patterns are large enough, this latter contribution is negligible. Unfortunately, random patterns are not self-supporting and, if forced to be so, undesirable correlations with first neighbors are introduced.

Array family	$\sum \mathbf{A}_{kl}\mathbf{G}_{ij}^2$	$\sum \mathbf{G}_{ij}^2$	SNR
(M)URA Random	$\frac{\rho N_T}{1-\rho}\left[\rho + \delta_{i-k,j-l}(1-2\rho)\right]$	$\frac{\rho N_T}{1-\rho}$	$\frac{\sqrt{N_T I_T}\sqrt{\rho(1-\rho)}(1-t)\psi_{ij}}{\sqrt{(1-t)\left[\rho + (1-2\rho)\psi_{ij}\right] + t + \xi}}$
Product arrays (PNP, MP, MM, NS)	N	N_T	$\frac{\sqrt{N_T I_T}\,\rho(1-t)\psi_{ij}}{\sqrt{(1-t)\rho + t + \xi}}$
Negative product arrays	$N_T \frac{\rho}{(1-\rho)^2}\left(\rho^2 - \frac{1}{N_T^2} + \frac{2}{N_T^2}N\delta\right)$	$\frac{N_T}{(1-\rho)^2}\left(\rho^2 + \frac{1-2\rho}{N_T^2}\right)$	$\frac{\rho(1-\rho)\sqrt{N_T I_T}(1-t)\psi_{ij}}{\sqrt{(1-t)\left(\rho^3 - \frac{\rho}{N_T^2} + 2\rho\frac{1-\rho}{N_T}\psi_{ij}\right) + (t+\xi)\left(\rho^2 + \frac{1-2\rho}{N_T^2}\right)}}$
NTHT (M)URA	$\begin{cases} 0 & \text{for } N_T^0(e^2-1) \text{ shifts} \\ N & \text{for } N_T^0 \text{ shifts} \end{cases}$	N_T^0	$\frac{\sqrt{N_T I_T}\sqrt{\rho/2}(1-t)\psi_{ij}}{\sqrt{(1-t)\rho + t + \xi}}$
Pinhole	$\delta(i-k, j-l)$	1	$\frac{\sqrt{I_T}(1-t)\psi_{ij}}{\sqrt{(1-t)\psi_{ij} + t + \xi}}$
Negative pinhole	$(N_T\text{-}1)[(N_T\text{-}2) + \delta(3\text{-}N_T)]$	$N_T^2 - 3N_T + 3$	$SNR_{ij} = \frac{\sqrt{\frac{N_T-1}{N_T-2}}I_T(1-t)\psi_{ij}}{\sqrt{(1-t)\left(1 - \frac{N_T-3}{N_T-2}\psi_{ij}\right) + (t+\xi)\frac{N_T^2-3N_T+3}{N_T^2-3N_T+2}}}$

Table 4.1: expressions needed for the calculation of the SNR via eq. (4.17) for the arrays discussed in this paper and final result. For NTHT (M)URAs N_T^0 is the total number of positions of the original array and $N_T = e^2\ N_T^0$ is still the total number of positions in the pattern. See text for e. Of course, $\mathbf{G} = \mathbf{G}^0$.

All results are gathered in Table 4.1, where the case of the negative product array is also included. It is not treated in detail because calculations are lengthy and do not offer particular insights. The result, however, will be later necessary to explain simulation results.

4.4 Comparing the SNR performance of different arrays

All masks except the random array offer ideal imaging but different SNRs. Given the number of parameters involved, a number of comparisons can be done. In general, the result is that coded apertures are favored for high-background and concentrated sources. We were particularly interested in two types of comparison. The first is between two apertures for different objects and background: a sample comparison between the pinhole and a NTHT mask based on a 50% open MURA is presented in detail. The second is a comparison of different apertures for a given object.

4.4.1 Effects of ψ and ξ on the SNR

From Table 4.1 the coded aperture has a SNR higher than a pinhole for $t = 0$ if:

$$(2 - N_T\rho)\xi \le N_T\rho\psi - 2\rho \tag{4.37}$$

Since the smallest URA is a 5×3 array with 8 open positions, the smallest open fraction is $\rho_{\min} = \frac{8}{N_T}$, so the left hand side is always negative:

$$\xi \ge \frac{\rho N_T}{2 - \rho N_T}\psi_{ij} - \frac{2\rho}{2 - \rho N_T} \tag{4.38}$$

This inequality can be interpreted as the separation of the (ψ, ξ) plane in two regions. The right hand side is the family of straight lines passing through the point $\left(\frac{2}{N_T}, 0\right)$. To each ρ corresponds a different line, which can be identified by its intersection with the ξ axis, i.e. $\left(0, \frac{2\rho}{\rho N_T - 2}\right)$. A coded aperture has a SNR higher than the pinhole depending on its parameters (i.e. N_T and ρ) and the image (i.e. ψ, and ξ). If the point of interest of the image lies above the line, the coded aperture is favorite.

We can already draw the conclusion that regardless of ξ, the coded aperture offers an advantage for all points such that $\psi_{ij} \geq 2 / N_T$. The first result of the SNR analysis is that coded apertures perform best for objects whose activity is concentrated in a few points, the best case being that of a point source. Around this first result a number of corollaries can be elaborated. For objects that fill the FoV, the average ψ_{ij} is $<\psi_{ij}> = 1/N_T$ and if background is low enough, there may be no advantage for points of average brightness. If the object fills only a fraction f^2 of the area of the FoV (f being the 1d reduction in the FoV), $<\psi_{ij}> = 1/f^2 N_T$, which shows that the smaller f, the higher the advantage over the pinhole. For $f < \sqrt{2}$ there surely is an advantage for the average source and if the object takes about 10% of the FoV (a 2×2 cm^2 area in a 9×9 cm^2 FoV), then the SNR advantage would be a factor of almost 7, which is a 49-fold reduction in time or activity at constant image quality.

The sources of a particular object can also be located on this graph. Since ξ is constant for all points, an object is represented by an horizontal line going from the minimum to the maximum ψ_{ij} in the object. In Figure 4.5 is presented the case of a thyroid case study (see section 0). First, the result that points with high ψ_{ij} are favored is found again: in fact, they are deeper into the coded aperture advantage region. A new result is that the SNR advantage of the coded aperture increases with background. In typical nuclear medicine conditions background is low, but this is not the case for CAFNA measurements, where background is responsible for the vast majority of recorded events.

Finally, in §4.3.3 the SNR of a NTHT (M)URA was proved to be an increasing function of ρ. This can be recognized in Figure 4.4 as well: for increasing ρ, the intersection of the boundary line with

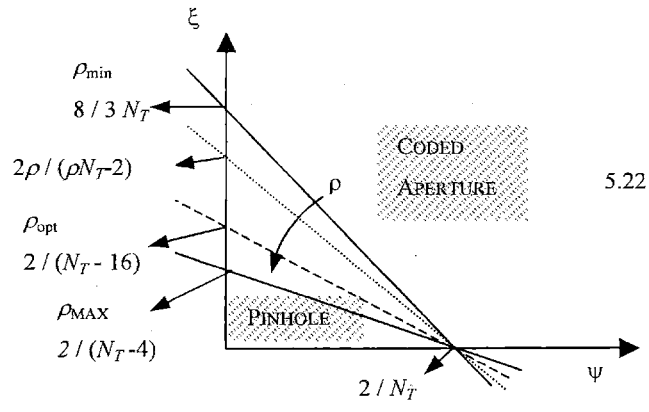


Figure 4.4: ψ-ξ plane indicating the situations for which a coded aperture offer a SNR higher than the pinhole.

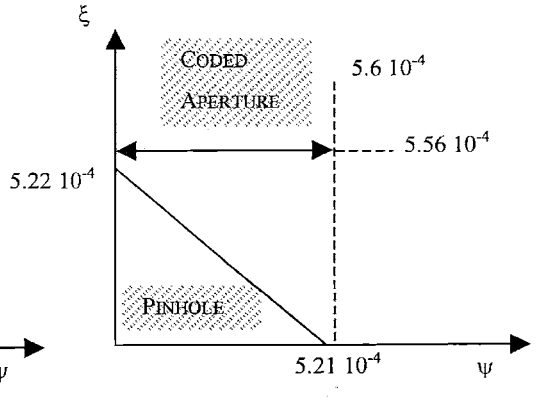


Figure 4.5: application to the thyroid case study. All image points are imaged with better SNR than they would have been with a pinhole.

the ξ axis occurs at ever lower points, expanding the region of better performance of the coded aperture. Since ρ must be greater than ρ_{min} but less than $\rho_{MAX} = 0.5$ (URA case), not all the lines of the family have a physical meaning. In Figure 4.4 the solid lines represent these limits. The line for $\rho = \rho_{opt} = 0.125$ is the one for which a real NTHT array gives best performance (which may or may not be better than that of the pinhole).

4.4.2 The SNR of different mask families

The graphs of Figure 4.8 were produced from the formulae of Table 4.1. A direct comparison of the SNR of different families can now be made by inspection. The following conclusions can be verified to be general with analytical methods too long and uninteresting to find place here.

For product and NTHT (M)URA arrays there is no optimum ρ for any combination of ξ and ψ_{ij}. For both families the SNR increases with ρ. For $\rho < 0.5$, a half-open (and even more an optimally open)

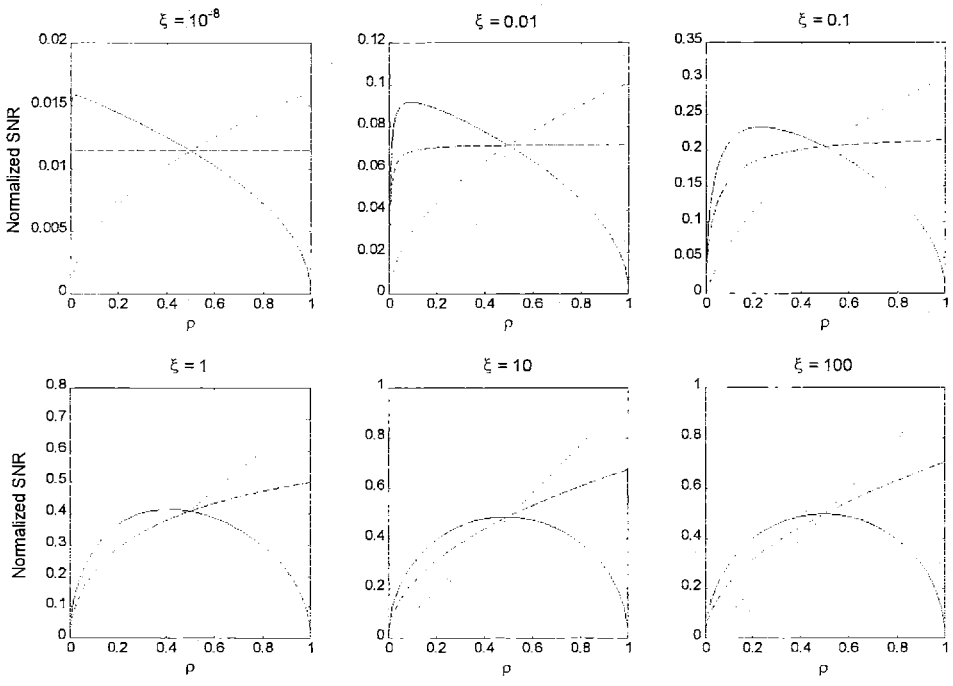


Figure 4.6: comparison of the SNR of different masks for zero transparency. All curves are drawn for $\psi_{ij} = 2.6 \times 10^{-4}$, the average value applicable to the thyroid study. The continuous line refers to URAs, the dashed to No-Two-Holes-Touching arrays and the dash-dot to product arrays, including the NS arrays of ref. [39]. The SNR is normalized to (i.e. it has been divided by) the SNR of the pinhole and $\sqrt{N_T}$ Multiply by $\sqrt{N_T}$ to obtain the advantage over the pinhole. Note that the SNR advantage increases with ξ.

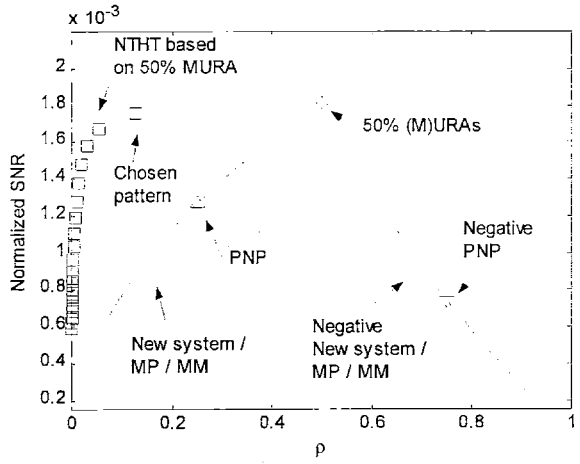


Figure 4.7: SNR curves for existing arrays. The SNR is normalized as in Figure 4.4. For this graph $\psi = 2.6 \times 10^{-4}$ and $\xi = 5.56 \times 10^{-4}$, as applies to the case study. The advantage over the pinhole for the NTHT mask of is about 1.5 in SNR terms (2.4 in time or activity) for this "average" image pixel.

URA always performs best. However, for $\rho > 0.5$ product arrays have a SNR higher than all other arrays. It is important to recall that the curves of Figure 4.8 are not concerned with the actual existence of patterns. They only state what the SNR of an array of a given family and open fraction would be if it existed. None of these families of arrays have patterns with an open fraction $\rho > 0.5$.

The observation that the negative of a good pair (**A**, **G**) is a good pair is encouraging because the negative of a low-open fraction mask must have a large open fraction, which makes it a good candidate for a high-SNR mask. Furthermore, the investigation of negative patterns is also relevant to near-field artifact correction methods (see §5.6). Both factors demanded the investigation of the negative of a product array. Unfortunately, it turns out that the SNR formula applicable to negative patterns is not the same as that of the original family (see Table 4.1) and the new calculation shows that the SNR for the negative array is much worse than predicted for the original array.

If the same curves of Figure 4.6 are sketched for existing arrays only, the situation is that of Figure 4.7, which is an explanation of our choice of the pattern family. The best SNR is achieved for a half-open (M)URA or *m*-sequence, but none of these patterns is self-supporting. A slightly inferior SNR is obtained with a NTHT (M)URA, which is self-supporting. Product arrays always offer a much lower SNR, this disadvantage increasing when the negative pattern, which is needed for near-field artifact removal, is also considered. This consideration does not affect (M)URAs, because the negative pattern has the same open fraction. This also holds for NTHT (M)URAs, because they can be considered (M)URAs with a hole smaller than the array position (§2.4.7). A more complete explanation of this

important result, which allows us to say that the negative of an NTHT (M)URA based on a 50% open pattern has still the same open fraction, and thus the same SNR, of the positive, is provided in Chapter 5.

4.5 Dependence of the variance on other sources

From the general expression for the SNR (eq. (4.17)), the SNR for one source depends on other sources through:

$$\sum_{k,l} \psi_{kl} \sum_{u,v} \mathbf{A}_{kl} \mathbf{G}_{ij}^2 \tag{4.39}$$

If $\sum_{u,v} \mathbf{A}_{kl} \mathbf{G}_{ij}^2$ were a constant over space, this term would be constant and the dependence would disappear. Gottesman and Schneid pointed out ([37]) that this is so for unimodular decoding arrays, i.e. those whose coefficients are all the same in absolute value. This is a desirable property of half-open (M)URAs and product arrays. One potential disadvantage of NTHT arrays is they do not satisfy this condition. To quantify the dependence of the SNR of a point on other sources of the object, the test proposed by Gottesman and Schneid was performed: given an object made of two point sources, one is kept fixed while the other is moved continuously from a pixel to a first neighbor. Let the shift be R. Measured in pixels, it ranges from 0, for no shift, to 1. In Figure 4.8 the variation of the SNR is evident, but it not as dramatic as for geometric arrays. This test is performed with point sources and no background, conditions that maximize the effect. In real cases "point sources" are spread over more than 1 pixel ([35]). For NTHT arrays, if the point source is a square of $e \times e$ pixels, no dependence at all is found

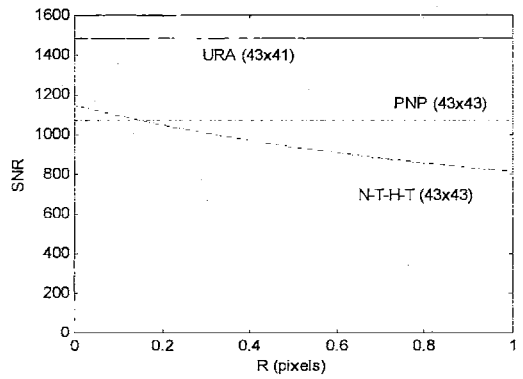


Figure 4.8: Dependence of the SNR on the position of sources. $I_T = 10^4$ and $\xi = 0$ as in Ref. [37]. For higher backgrounds the dependence is not as strong, but this is not the case in Nuclear Medicine applications.

because $\sum_{k,l}\psi_{kl}\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2$ is constant. In fact, from Table 4.1 $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2$ is 0 for shifts that superimpose blank lines of **G** with holes of **A** and N otherwise. This means that $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2$ is a matrix of zeros with peaks of height N arranged on a square grid and spaced by e in each direction. If the source is uniformly spread over a few pixels as mentioned, this array must be convolved with the $e \times e$ square, giving a constant. Indeed, we found in the image structures due to this dependence on the SNR only for a particular source used in a preliminary experiment. These are the capillaries of Figure 6.7 and Figure 6.8. In these cases, points with no signal are still affected by noise through the cross-talk term (4.39). Since the sources are not spread over a $e \times e$ square, but along a line, the peaks of the function $\sum_{u,v}\mathbf{A}_{kl}\mathbf{G}_{ij}^2$ are replicated by ψ_{ij} only in directions parallel to those of the capillaries, leaving intermediate regions empty. The effect is to create a striped background, as can be verified in the above-mentioned figures. A broader theoretical discussion of noise correlations in quantitative measurements can be found in ref. [51], which, however, does not address directly the case of imaging.

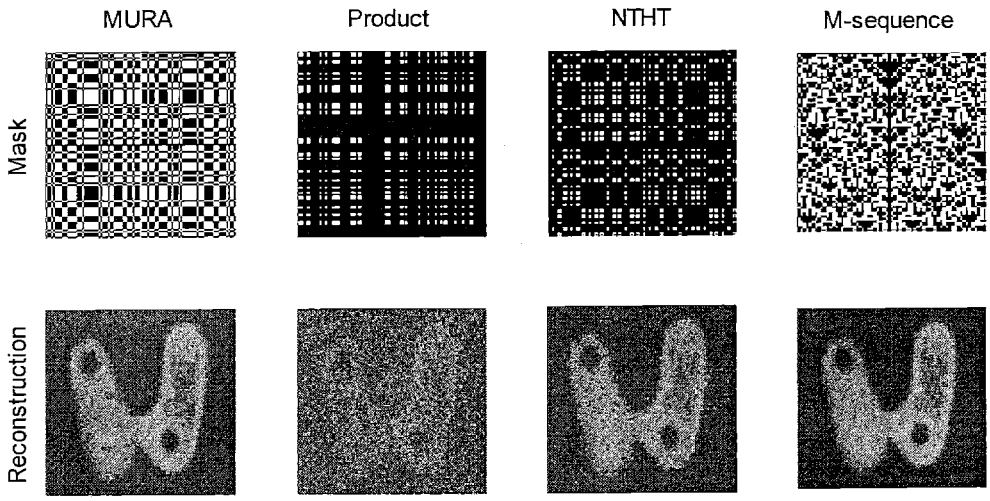


Figure 4.9: coded apertures used for thyroid phantom simulations. Top row: 79 × 79 MURA. Open fraction: 0.5. Mask pixel size: 0.921 mm, geometric resolution: 1.139 mm. Product: 77×77 NS array built from a 7 and an 11 1d URA. Open fraction: 0.097. Mask pixel size 0.9 mm, geometric resolution: 1.2 mm. NTHT array based on a 31×31 MURA, with $e = 2$. Open fraction: 0.125. Mask pixel size 1.1 mm, geometric resolution: 1.5 mm. 63 × 65 *m*-sequence. Open fraction: 0.5. Mask pixel size 1.1 mm, geometric resolution: 0.14 mm. For all masks: FoV 9 × 9 cm, thickness 1.5 mm (1% penetration at 140 keV). Only the elementary pattern of a 2×2 mosaic is shown. Bottom row: simulation results with near-field artifacts suppression.

4.6 Simulations

Several simulations were run to verify theoretical predictions. For purposes of comparison with the literature and previous results obtained by our group, the object simulated was a thyroid phantom. The simulation was a 17.5-min exposure with 40 cm object-to-detector distance. The activity was 200 μCi of ^{99m}Tc (140 keV). The mask designs under investigation are shown in Figure 4.9. They were chosen to provide the maximum resolution for a field of view of 9 × 9 cm given the constraint that the projection on the detector should take no more than 160 × 160 pixels with $\alpha = 2$. Given the extension of the object and the fact that the open fraction of the NS array was chosen as close as possible to the optimum value prescribed in ref. [52], according to ref [39], we should have expected best performance for this aperture.

The simulation results are also shown in Figure 4.9, after near-field artifact suppression (Chapter 5), which seems equally good in all cases. In agreement with theory, best performance is shared by *m*-sequence and MURA. Slightly worse is the NTHT pattern, which still shows better SNR than the product array, despite the higher open fraction. The error of ref [39] (also found in [3] and [40]) was to extend the results of ref. [52], derived for and, thus, applicable only to URAs, to product arrays. A hint might have been that the SNR of PNP arrays, which are strictly related to the NS arrays, was known, under particular conditions (no mask transparency and zero background), to compare unfavorably with URAs ([37]).

4.7 Observation on the relation between SNR and sensitivity

Specifications of collimators and pinholes do not list their SNR. An indirect measure is offered in terms of sensitivity, which is by definition the number of counts obtained per unit time and activity in the source. This is indeed a good indirect measure of the SNR for such non-multiplexing devices. Since there are no superpositions to be undone, every count is a "good" count for the system. In fact, from the expression of the SNR for the pinhole, for no transmission and background one gets:

$$SNR_{ij} = \sqrt{I_T \Psi_{ij}} = \sqrt{\mathbf{O}_{ij}} \tag{4.40}$$

which means that the SNR of a source in the object is the square root of the number of counts coming from that source, i.e. it is the square root of the sensitivity of the system multiplied by activity and time.

The relationship between sensitivity and SNR is not as direct for coded apertures, because of decoding. In fact, for an (M)URA, for no transmission and background:

$$SNR_{ij} = \sqrt{\rho N_T I_T}\,\psi_{ij} = \sqrt{N I_T}\,\psi_{ij} \tag{4.41}$$

but the sensitivity to a source is proportional to $N I_T \psi_{ij}$. This means that sensitivity is subject to a $\sqrt{\psi_{ij}}$ loss before it is translated to SNR. For a point source there is no loss, confirming that the origin of the loss is in the superpositions. In this sense not all counts, but only a fraction $\sqrt{\psi_{ij}}$ of them, are "good" counts to a coded aperture and the tremendous advantages in sensitivity seen in the next Chapters should not be overestimated.

Chapter 5 ARTIFACT THEORY

Even a device as simple as the pinhole camera is affected by non-idealities. A perfect pinhole is a dimensionless hole in an infinitely thin, but yet perfectly opaque, screen. In these conditions, resolution and the field of view are limited by the detector. Intensities are modulated by a $\cos^3(\theta)$ factor, but with knowledge of the object-to-detector distance the collected data can be corrected and the effect compensated. In reality, the hole must have a finite size, which limits resolution, and thickness, which cuts rays entering at angles, limiting the field of view and changing the brightness of off-axis sources, which becomes a function of the incidence angle even in far-field geometry. Finally, noise in the data make compensation for near-field geometry only approximate.

Coded apertures suffer from the same limitations. The most impressive artifacts, however, come from the $\cos^3(\theta)$ factor. With it, the projection of the mask does not look exactly like the mask itself. When a correlation is taken to decode the data, **G** is not correlated with **A**, but with **A** modulated by $\cos^3(\theta)$. The result is not a δ and artifacts appear in the image. The scope of this Chapter is to gain insight in the process. A mathematical framework is developed to predict artifacts and develop solutions. Predictions are compared to simulation, experiment and literature data. Sampling effects, such as border and non-integer α are also considered. The last section is dedicated to mask thickness.

The only paper discussing artifacts we know is ref. [54], where mask transmission, geometry of the open elements, finite positional resolution of the detector, its possible malfunctions, alignment problems and sources in the PCFV are treated.

5.1 Mask transmission

An ideal mask is infinitely thin and perfectly opaque, i.e. blocks all γ-rays falling on opaque positions. In reality, these rays have a finite probability t of passing through the mask. From the theory of γ-ray attenuation, the fraction of rays passing through opaque locations (t, transmission or transparency of the mask) is:

$$t = e^{-\mu\rho x} \tag{5.1}$$

where ρ is the density of the mask material, μ its attenuation coefficient, evaluated at the energy of interest, and x the mask thickness. For example, the attenuation coefficient of tungsten at 140 keV, the energy of ^{99m}Tc γ-rays, is 1.76 cm^2/g. Since the density of the material is 19.3 g/cm^3, from eq. (5.1), for a 1.5-mm-thick mask $t = 0.006$ or 6‰. For ^{111}In (171, 245 keV) the attenuation coefficient in tungsten (ρ = 19.3 g/cm^3) is 0.458 cm^2/g:, which makes the same mask 26% transparent.

Eq (1.1) does not hold and must be corrected. When transmission is modeled, the pattern recorded at the detector is the contribution of a fraction (1-t) of photons that "see" an ideal mask (and which are, then, described by eq. (1.1)), plus a fraction t of photons that do not see the mask at all:

$$\mathbf{R} = (1-t)\,\mathbf{O} \times \mathbf{A} + t\,\mathbf{O} \times \mathbf{1} = (1-t)\,\mathbf{O} \times \mathbf{A} + \text{constant} \tag{5.2}$$

where **1** is an entirely open mask and the result of Appendix A.2 was applied. Mask transmission reduces the signal in the ideal image and adds background. These effects do cause a loss in SNR, as discussed in Chapter 4, but do not distort the image or contribute artifacts. Ref. [54] reports experimental verifications of this argument and adds a SNR formula extending the work of Fenimore along the lines seen in Chapter 4.

5.2 Sampling

By sampling artifacts we refer to artifacts caused by the fact that the detector is not a smooth continuum, but is, effectively (but not physically), made of pixels. They are divided in border and non-integer α artifacts.

5.2.1 Border

In the ideal case, if α, the number of pixels on which is projected the shadow of a mask position, is an integer, the projection of a point source covers exactly a square of α × α pixels. All pixels in the square are completely illuminated and the reconstruction is exactly as expected for an ideally continuous detector. In reality, this happens only for very particular positions of the point source in the object: Figure 3.1 shows the example of a projection not matching detector pixels but covering a part of an α+1 × α+1 square. Of these, pixels on the border of the square are only partially illuminated. If α is integer, the distribution of γ-rays on different pixels of the square is the same for all mask holes, and simply takes place at different parts on the detector. Furthermore, if $\alpha \geq 2$ at least one pixel for every mask hole is

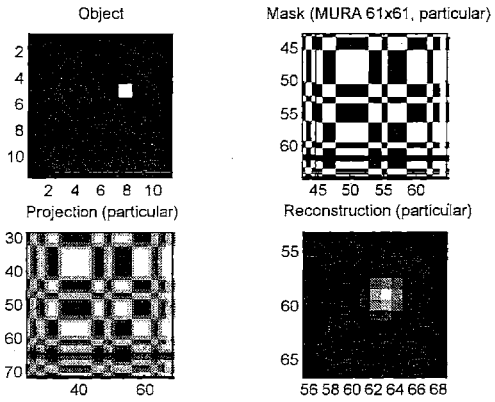


Figure 5.1: border effect on the reconstruction of a point source. In the reconstruction note the gray pixels surrounding the white center pixel. These arise from the gray areas surrounding the black and white zones of the mask projection: these are detector pixels partially illuminated because of border effect.

totally illuminated. These pixels are spaced by α pixels, on a rectangular grid, in both directions. In the case of δ decoding (see section 2.7), at the corresponding reconstruction positions the brightness of the source is reconstructed as if the shadow was cast matching exactly the boundary of the pixels. This is true for all totally illuminated pixels. Border pixels, however, show a reduced brightness. However, if a certain pixel has a brightness reduced by a certain factor, all pixels shifted by α positions from it, also have the same reduction. Therefore, next to the main peak, appears another perfect reconstruction of lower brightness. In conclusion, a point source which is ideally reconstructed on a $\alpha \times \alpha$ square of pixels of constant brightness is in reality reconstructed on a α-1 $\times$ α-1 square of pixels of constant brightness surrounded by a border of pixels of lower brightness. This is what we called the border effect.

Of course, the total activity is the same, but is distributed differently in the two cases. It is important to notice that the height of the peak does not change with source shift if $\alpha \geq 2$, because at least one totally illuminated pixel is always present. This is not so for $\alpha < 2$. For example, for $\alpha = 1$, depending on the position of the source, the activity can be distributed on a pixel only, which gives a peak with a height of, say, 1, with neighbors at 0, or on 4 pixels, each with a height of ¼. In between these limits a number of combinations is possible, all giving different maximum heights. This can be very confusing when two or more sources are present in the image because sources of the same brightness can result in different maximum heights, and thus be visualized by widely different colors, giving the false impression of different activity. To avoid the problem, we always chose $\alpha \geq 2$, even if in literature this assumption is deemed "unnecessarily conservative" ([33]).

Ref. [54] indicates that these artifacts can be deconvolved. The approach, in our opinion, has two limitations. First, noise may be a limiting factor. Second, the PSF to deconvolve depends on the position of the point source in the FoV and is different for different point sources.

5.2.2 Non-integer α

Much more serious is the effect of a non-integer value of α. For a given mask and detector, α depends on distances from the mask:

$$\alpha = \frac{m\, p_m}{p_d} \tag{5.3}$$

For non-integer values of α it is impossible to scale the decoding pattern, which is an array stored in the computer, to match exactly the size of the projection of the mask. This implies that **G** and **A** can not have the same size. The effects are very difficult to predict analytically and were simulated. In Figure 5.2 is shown the effect of decoding for α = 2 point sources projected for α only approximately equal to 2. A 5% difference in α is already enough to disrupt the PSF. Unfortunately, complex and orderly structures appear. Artifacts are expected to contribute to the reconstructed images. This is also a serious blow to hopes for laminography (see section 8.4), because it proves that off-focus planes do not contribute a uniform, blurred background.

These reconstructions are not equivalent to a 3d PSF because here different data are decoded at the same depth, which, in principle, is not the same as decoding the same data at different depths. This is

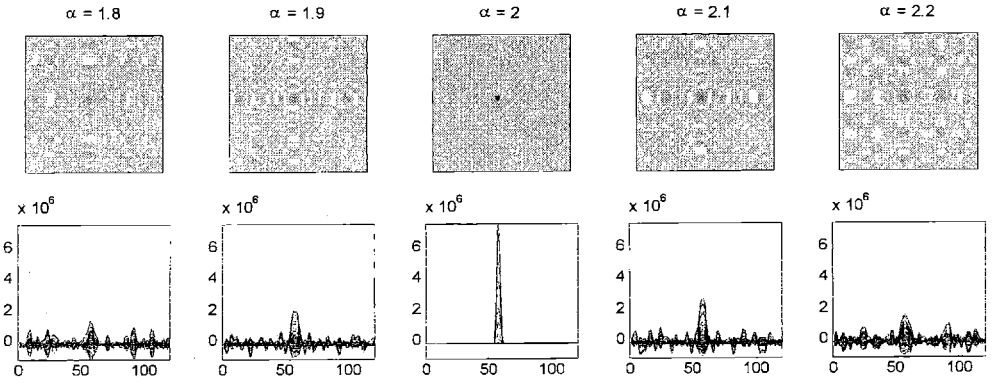


Figure 5.2: decoding for α = 2 of a point source projected for different values of α. 62 × 62 NTHT MURA was used with very high activity source in a far-field geometry to avoid interference with statistical noise and near-field artifacts. The PSF is ideal only for exact decoding.

indeed possible and in a continuous way, i.e. it is possible to decode the data with non-integer α, and is explained in Chapter 8.

Two solutions are possible to avoid these artifacts. The first is to focus the coded aperture camera very carefully. In our experience this is best done by trial and error, with the help of the strategy outlined in section 2.8. The second is to be able to decode for any value of α. In either case artifacts are completely avoided for 2d or very thin 3d objects only. A measure of how thin is an object can be obtained solving eq. (2.107) for $d\alpha$:

$$\frac{d\alpha}{\alpha} = \pm\frac{b}{za}\frac{da}{2} = \pm\frac{m-1}{z}\frac{da}{2} \tag{5.4}$$

where da is the thickness of the object and the factor ½ comes from the assumption that the on-focus plane is at the center of the object. If $d\alpha / \alpha > 5\%$, Figure 6.1 shows that artifacts from the external object planes should be expected. Of course, the limit on the maximum acceptable $d\alpha$ is largely arbitrary.

5.3 Rotational misalignment

Artifacts arise if the projection of the mask is not parallel to detector pixels but is rotated at some angle. These artifacts are also complex. We did not undertake a full investigation of the issue because proper alignment can be reached by trial and error using a point source and verifying the position of the projection. To allow fine tuning, it is important that any flange connecting the mask to the detector have a rotational degree of freedom to allow the fine tuning necessary.

5.4 Near-field artifacts

From geometric optics, the expression giving the number of counts recorded at the detector position $\vec{r}_i$ is:

$$\mathbf{R}(\vec{r}_i) \propto \int_{\vec{r}_o} \mathbf{O}(\vec{r}_o)\mathbf{A}\left(\frac{a}{z}\vec{r}_i + \frac{b}{z}\vec{r}_o\right)\cos^3(\theta)d^2\vec{r}_o \tag{5.5}$$

where $\theta = \text{atan}\left(\left|\vec{r}_i - \vec{r}_o\right|/z\right)$, and all other symbols were defined in section 2.2. Defining:

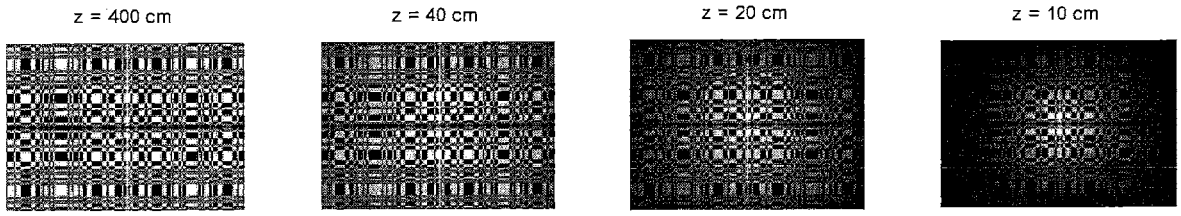


Figure 5.3: $\cos^3(\theta)$ modulation as a function of object-to-detector distance. Infinitely thin mask and activity to avoid confusion with mask thickness and noise effects.

$$\vec{\xi} = -\frac{b}{a}\vec{r}_o,\ \mathbf{O}'(\vec{r}) = \mathbf{O}\left(-\frac{a}{b}\vec{r}\right),\ \text{and}\ \mathbf{A}'(\vec{r}) = \mathbf{A}\left(\frac{a}{z}\vec{r}\right) \tag{5.6}$$

the form:

$$\mathbf{R}(\vec{r}_i) \propto \iint_{\vec{\xi}} \mathbf{O}'(\vec{\xi})\mathbf{A}'(\vec{r}_i - \vec{\xi})\cos^3\left[\operatorname{atan}\left(\frac{\left|\vec{r}_i + \frac{a}{b}\vec{\xi}\right|}{z}\right)\right] d^2\vec{\xi} \tag{5.7}$$

is reached. Recall from section 2.2 that $\mathbf{O}'$ and $\mathbf{A}'$ are, respectively, a rescaled and reflected form of the object and a rescaled version of the mask pattern. Eq. (5.7) is not in the form of a convolution because of the near-field term $\cos^3(\theta)$ whose effect on the projection is shown in Figure 5.3: it is not a modulation of the object by a constant factor $\cos^2(\theta_0)$, as for off-axis sources in far-field problems ([33]).

From coded aperture theory we know that the reconstructed image $\hat{\mathbf{O}}$ is obtained from $\mathbf{R}$ via periodic correlation with the decoding pattern $\mathbf{G}$ associated with the mask $\mathbf{A}$:

$$\hat{\mathbf{O}} = \mathbf{R} \otimes \mathbf{G} \tag{5.8}$$

If no assumptions are made, this form is not analytically tractable.

5.4.1 The far-field approximation

If the object is far from the detector so that $\left|\vec{r}_i - \vec{r}_o\right| << z,\ \forall\left(\vec{r}_i, \vec{r}_o\right)$ then:

$$\cos^3(\theta) \cong 1 \tag{5.9}$$

This is the far-field approximation. Under this hypothesis, eq. (5.7) is reduced to the convolution:

$$\mathbf{R} = \mathbf{O}' * \mathbf{A}' \tag{5.10}$$

which is the far-field case already discussed treated at length and summarized here for convenience. For ideal pairs (**A**, **G**), eq. (5.8) becomes:

$$\hat{\mathbf{O}} = (\mathbf{O}' * \mathbf{A}') \otimes \mathbf{G} = \Re[\mathbf{O}' * (\mathbf{G} \otimes \mathbf{A}')] = \Re(\mathbf{O}' * \delta) = \Re(\mathbf{O}') \tag{5.11}$$

where $\Re$ is the reflection operator (see eq. (2.36)). The reconstructed object is the object itself apart from a rescaling constant. However, if condition (5.9) does not hold, eq. (5.7) does not assume the simplified form of eq. (5.10) and this result is not reached; the image is corrupted in some way.

5.4.2 The near-field case

In some applications it is imperative to collect data as close to the source as possible. For instance, in Nuclear Medicine sensitivity must be maximized to keep the dose to the patient as low as reasonably achievable. Hence, in such applications, $|\vec{r}_i - \vec{r}_o|$ is often comparable to z and eq. (5.9) does not hold. To make the problem still mathematically treatable we expanded the near-field term of eq. (5.7) in Taylor series to the second order with center $\vec{r}_i$. The result is[13]:

$$\cos^3\left(\operatorname{atan}\left(\frac{\left|\vec{r}_i + \frac{a}{b}\vec{\xi}\right|}{z}\right)\right) \cong \cos^3\left(\operatorname{atan}\left(\frac{|\vec{r}_i|}{z}\right)\right)\left\{1 - \frac{3/z^2}{\left(1+\frac{|\vec{r}_i|^2}{z^2}\right)}\left[\vec{r}_i \circ \frac{a}{b}\vec{\xi} + \frac{1}{2}\frac{a}{b}\left|\vec{\xi}\right|^2 - \frac{5/2z^2}{\left(1+\frac{|\vec{r}_i|^2}{z^2}\right)}\left(\vec{r}_i \circ \frac{a}{b}\vec{\xi}\right)^2\right]\right\} \tag{5.12}$$

where $\circ$ indicates scalar product. This expansion is the more accurate the larger the margin by which the condition:

$$\frac{\left|\frac{a}{b}\vec{\xi}\right|}{|\vec{r}_i|} = \frac{|\vec{r}_o|}{|\vec{r}_i|} < 1 \tag{5.13}$$

---

[13] Before starting this calculation it is very convenient to use: $\cos(\arctan(x)) = \dfrac{1}{\sqrt{1+x^2}}$

is true. When high resolution is sought, magnification tends to be high, so a/b is generally small. Another way of looking at the same condition is to recognize that $\vec{r}_o$ is a variable spanning the object. Because of high magnification, if the object fits in the field of view, it is typically much smaller than the detector and $\vec{r}_o < \vec{r}_i$, which makes the Taylor approximation a good one over most of the detector. In our applications stopping at second order was sufficient to explain the artifacts we were seeing. Note that high magnification is only a sufficient condition for eq. (5.13) to hold, but it can be satisfied in many other cases.

Eq. (5.12) breaks the near-field term $\cos^3(\theta)$ into the sum of a zero, first and second order contribution: eq. (5.7) is decomposed to a sum of parts that can be examined one at a time.

5.4.3 Artifact prediction

In the following, an attempt to predict the shape of near-field artifacts in the reconstructed images is made. Examples will be taken from the same thyroid phantom study of section 4.6. Images were taken for an object-to-detector distance of 40 cm and a magnification factor $m = 4.3$. The mask used was a 62 × 62 NTHT based on a 31×31 MURA and was made of 1.5-mm-thick tungsten. Its pinholes were 1.1-mm-wide.

Zero order correction

If the expansion is stopped at zero order, from eq. (5.12) one has:

$$\cos^3(\text{atan}(\frac{\left|\vec{r}_i + \frac{a}{b}\vec{\xi}\right|}{z})) \cong \cos^3(\text{atan}(\frac{|\vec{r}_i|}{z})) \tag{5.14}$$

which does not depend on $\vec{\xi}$ and substituted in eq. (5.7), gives:

$$\mathbf{R}(\vec{r}_i) = \cos^3(\text{atan}(\frac{|\vec{r}_i|}{z})) \iint_{\vec{\xi}} \mathbf{O}'(\vec{\xi})\, \mathbf{A}'(\vec{r}_i - \vec{\xi})\, d^2\vec{\xi} \tag{5.15}$$

Since $\mathbf{R}$ is not the convolution of $\mathbf{O}'$ and $\mathbf{A}'$, correlation with $\mathbf{G}$ does not produce the object. However, the near-field effect is reduced to a prefactor depending on the detector coordinate $\vec{r}_i$ only. It is, thus, easy to correct zero order artifacts exactly by dividing the projection data $\mathbf{R}$ by this prefactor. The projection is now a convolution and we are reduced to the far-field case of §5.4.1. A physical

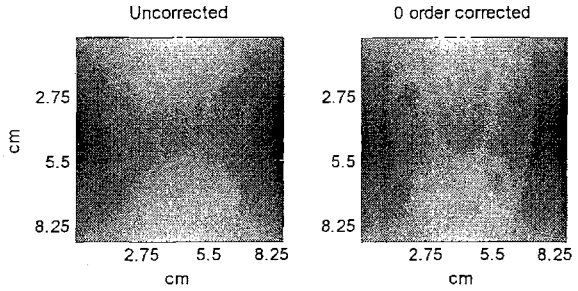


Figure 5.4: effect of zero order correction. The white rims surrounding the cold spots and the right lobe of the thyroid are not artifacts but part of the simulated object (see Figure 7.12).

interpretation is that, in this approximation, artifacts are the same as if the object were all concentrated at the center of the field of view. So, if the object *is* a point source at the center of the field of view, the correction restores the ideal image.

All decoding programs include this correction. The effects can be seen in Figure 5.4. Even if some improvement is achieved, the artifacts, the bright "bows" seen at the top and bottom, are hardly eliminated. Higher order terms are not negligible and must be analyzed. Also note that the noise level of the image in Figure 5.4 is somewhat increased. By dividing by a constant we are artificially increasing the number of counts of side pixels. The average value is restored, but the variance remains that of a lower number of counts.

First order: centering the mask pattern and the object

From eq. (5.12) the expression for first order term is:

$$\cos^3\left(\operatorname{atan}\left(\frac{\left|\vec{r}_i+\frac{a}{b}\vec{\xi}\right|}{z}\right)\right)\Bigg|_{\mathrm{I}} \cong -\cos^3\left(\operatorname{atan}\left(\frac{|\vec{r}_i|}{z}\right)\right)\frac{3/z^2}{\left(1+\frac{|\vec{r}_i|^2}{z^2}\right)}\left[\vec{r}_i \circ \frac{a}{b}\vec{\xi}\right] \tag{5.16}$$

where the bar at the left-hand side indicates that this expression includes the first order term only. Substitution in eq. (5.7) leads, after zero order correction, to:

$$\mathbf{R}(\vec{r}_i) \propto \frac{\vec{r}_i}{z^2+|\vec{r}_i|^2} \circ \iint_{\vec{\xi}} \vec{\xi}\mathbf{O}'(\vec{\xi})\mathbf{A}'(\vec{r}_i-\vec{\xi})\; d^2\vec{\xi} \tag{5.17}$$

To reach a useful interpretation, it is important to bear in mind that $\mathbf{A}'$ is a function describing the aperture. Typical coded apertures are binary, i.e. can assume only two values, 0 and 1, indicating, respectively, the closed and open positions of the mask. In light of this, one can recognize that the integrand is the center of mass of the object "cut" by $\mathbf{A}'$. Convolution makes the result a function of the shift of $\mathbf{A}'$. Now, if $\mathbf{A}'$ covers the field of view uniformly and the object is also reasonably uniform, the result is not a strong function of shift and gives an approximately constant contribution. In the thyroid case study, the integral in eq. (5.17) gives **R** only a small modulation of the main structure coming from the first factor. Consequently, the integral can be replaced with a constant vector, $\vec{O}'$. The result can then be substituted into eq. (5.7). Since $\vec{O}'$ is constant, it can be taken out of the correlation integral. The result is:

$$\hat{\mathbf{O}} \propto \vec{O}' \circ \iint_{\vec{r}_i} \frac{\vec{r}_i}{z^2 + |\vec{r}_i|^2} \mathbf{G}(\vec{r}_i + \vec{\eta})\, d^2\vec{r}_i \tag{5.18}$$

The important consequence here is that, under the above-mentioned hypotheses, the shape of first order artifacts depends on **G** only and can be calculated with this integral. This noted, a further

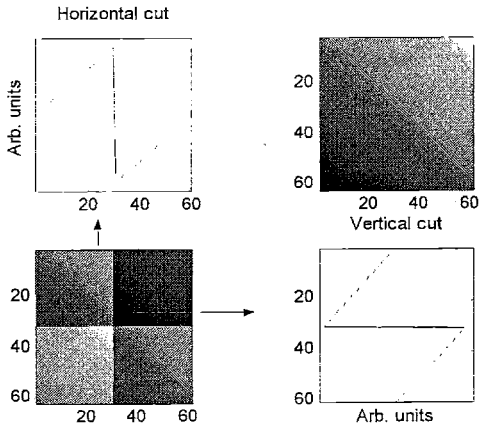


Figure 5.5: center of mass of the decoding pattern **G** as a function of shift for a MURA. The expected form of the artifact is at the bottom left. Cross sections of this function are shown on its top and right. If a shift different than that in which the arrays are given by generation rules had been used, the result would have been that at the top right. This shift corresponds to putting the solid black row and the solid white column of the MURA at the center of the mask.

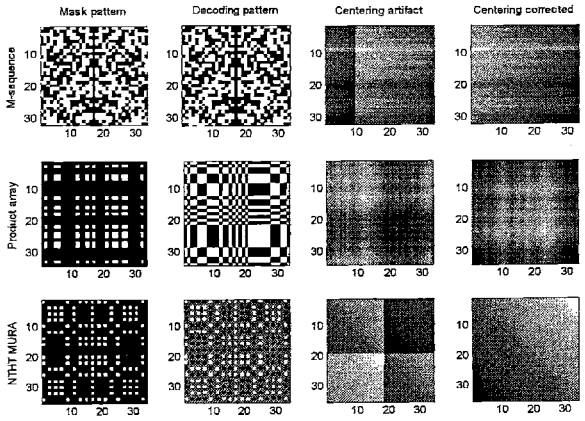


Figure 5.6: first order artifacts for three different array families. Top: 31×33 *m*-sequence. Middle: 33×33 Pseudo-Noise-Product (PNP). Bottom: 34×34 NTHT MURA. Masks and decoding arrays are shown in the pattern centered shift. Non corrected artifacts are shown for the pattern as obtained from the generation rule found in literature. The correction is substantial for *m*-sequences and NTHT patterns, while it is of dubious effectiveness for product arrays.

approximation allows the discussion of this same result in more intuitive terms. However, the following observations could have been referred to this more involved form as well.

In the thyroid case study, $\vec{r}_i$ spans at most a 38.6 × 38.6 cm^2 area and z is 40 cm. For these values, substitution of the fraction in the integral with $\vec{r}_i/z^2$ is accurate within 28% at the worst points:

$$\hat{\mathbf{O}} \propto \vec{O}' \circ \iint_{\vec{r}_i} \vec{r}_i \mathbf{G}\left(\vec{r}_i + \vec{\eta}\right) d^2\vec{r}_i \tag{5.19}$$

Both factors are important in the discussion of first order artifacts. The integral, just like eq. (5.17) for $\mathbf{O}'$, gives the position of the center of mass of the open positions of the decoding pattern **G** as a function of decoding position (which is the same as the shift of **G**). Therefore, the object reconstruction $\hat{\mathbf{O}}$ involves an additive term depending on the particular form of **G**. The case of a MURA pattern is shown in Figure 5.5. The image at the bottom left, shows that this function has a sudden drop at its center, resulting in a vertical and a horizontal line at the center of the reconstruction. If **G** (and thus the mask pattern **A**) is not taken as generated by the rules found in literature but is shifted by half a period in both directions, these drops can be moved to the borders of the image, removing the most unpleasant part of the artifact. We called this technique pattern centering. The case of other patterns (a 31×33 *m*-sequence, a 33×33 product array and a NTHT MURA is shown in Figure 5.6.

The first factor in eq. (5.19), the center of mass of the object, can be used to remove this artifact completely. In fact, if the object were centered on the field of view, $\vec{O}'$ would be zero, canceling the term. From a practical point of view one can take a first, raw, image to estimate $\vec{O}'$ and then make the necessary adjustments before taking a second picture. Examples are presented in §5.5.1. However, we shall see that taking two pictures is not necessary.

Second order: mask and anti-mask

If the object is centered on the FoV, first order artifacts disappear, but second order artifacts, normally hidden by the stronger lower order ones, become visible. This must be the case of Figure 5.4, because the terms so far analyzed still do not explain the "bows" corrupting the image. The starting point is the substitution in eq. (5.7) of the second order terms of eq. (5.12):

$$\cos^3(\text{atan}(\frac{\left|\vec{r}_i+\frac{a}{b}\vec{\xi}\right|}{z}))\Big|_{II} \cong \cos^3(\text{atan}(\frac{|\vec{r}_i|}{z}))\frac{3/z^2}{\left(1+\frac{|\vec{r}_i|^2}{z^2}\right)}[\tfrac{1}{2}\tfrac{a}{b}|\vec{\xi}|^2-\frac{5/2z^2}{\left(1+\frac{|\vec{r}_i|^2}{z^2}\right)}(\vec{r}_i\circ\tfrac{a}{b}\vec{\xi})^2] \tag{5.20}$$

This time two terms must be considered. The first is:

$$\mathbf{R}(\vec{r}_i)\propto\frac{1}{z^2+|\vec{r}_i|^2}\iint_{\vec{\xi}}\mathbf{O}'(\vec{\xi})\mathbf{A}'(\vec{r}_i-\vec{\xi})|\vec{\xi}|^2d^2\vec{\xi} \tag{5.21}$$

The integrand is the second moment of inertia of $\mathbf{O}'$, cut by $\mathbf{A}'$, with respect to an axis perpendicular to the object plane and passing through its center. If the open positions of $\mathbf{A}'$ are uniformly distributed (which is the case for all patterns of this paper) the shift of $\mathbf{A}'$ does not greatly modify the result. The integral is, then, approximately constant. If we note, as we did for first order artifacts, that $|\vec{r}_i|^2<z^2$, the factor outside the integral does not depend strongly on $\vec{r}_i$. Therefore, upon decoding, $\mathbf{G}$ is convolved with a constant, giving a constant term that can be neglected.

The second term of eq. (5.20) gives:

$$\mathbf{R}(\vec{r}_i)\propto\frac{|\vec{r}_i|^2}{\left(z^2+|\vec{r}_i|^2\right)^2}\iint_{\vec{\xi}}\mathbf{O}'(\vec{\xi})\mathbf{A}'(\vec{r}_i-\vec{\xi})(\underline{r}_i\circ\vec{\xi})^2d^2\vec{\xi} \tag{5.22}$$

where $\underline{r}_i$ is the unit vector having the same direction as $\vec{r}_i$. The integrand is the moment of inertia of $\mathbf{O}'$, cut by $\mathbf{A}'$, with respect to the axis in the plane of the object perpendicular to $\underline{r}_i$ as a function of the shift of $\mathbf{A}'$. In the usual assumption that the open positions of $\mathbf{A}'$ are uniformly distributed, it is a function with little dependence on $|\vec{r}_i|$. The integrand can be approximated with $\rho I(\underline{r}_i)$, where ρ is the open fraction of $\mathbf{A}$ and I the moment of $\mathbf{O}'$ we have just discussed:

$$I(\underline{r}_i)=\iint_{\vec{\xi}}\mathbf{O}'(\vec{\xi})(\underline{r}_i\circ\vec{\xi})^2d^2\vec{\xi} \tag{5.23}$$

The contribution to the decoded image is obtained by substitution in eq. (5.7):

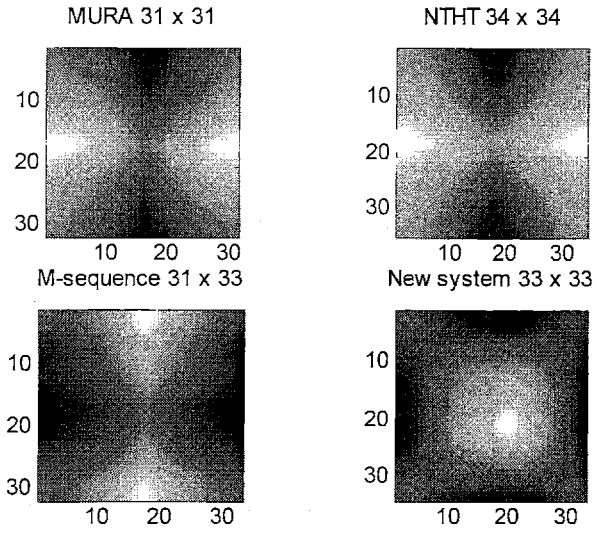


Figure 5.7: second order artifacts for the arrays of Figure 5.5 and Figure 5.6 (after pattern centering) calculated with eq. (5.25). Note the similarity of MURAs and NTHT MURAs, which effectively are (M)URAs with smaller holes. This is also true of first order effects.

$$\hat{\mathbf{O}} \propto \iint_{\vec{r}_i} \frac{|\vec{r}_i|^2}{\left(z^2+|\vec{r}_i|^2\right)^2} \rho I(\underline{r}_i)\mathbf{G}(\vec{r}_i+\vec{\eta})\, d^2\vec{r}_i \tag{5.24}$$

Dependence on the object enters this equation only through its moment of inertia I. Therefore it is stronger the less isotropic the object; as, for example, a line source. For isotropic objects, like a circle, I is a constant ($\pi R^4 / 4$ where R is the radius of the circle) and comes out of the integral. In this case, artifacts depend again on **G** only, i.e. the mask family we are using:

$$\hat{\mathbf{O}} \propto \rho I \iint_{\vec{r}_i} \frac{|\vec{r}_i|^2}{\left(z^2+|\vec{r}_i|^2\right)^2} \mathbf{G}(\vec{r}_i+\vec{\eta})\, d^2\vec{r}_i \tag{5.25}$$

As a first approximation, one can ignore the denominator and conclude that second order artifacts have the shape of the moment of inertia of **G** (with respect to an axis perpendicular to its plane and passing through its the center) as a function of shift. In the calculations, however, one can implement as easily the whole eq. (5.25): the application to different array families is in Figure 5.7.

Not surprisingly, at second order the second moments of **O** and **G** appear. For isotropic objects, artifacts depend only on a quantity we know (**G**) and can be predicted independently of the object.

However, we found empirically that the prediction is accurate even for objects that, as our test object, do not seem particularly isotropic.

Mask transmission in near field artifacts

The argument of section 5.1 can be repeated to see if mask transmission generates artifacts in near field. The starting point is rewriting eq. (5.7) as:

$$\mathbf{R}(\vec{r}_i) \propto (1-t)\int_{\vec{r}_o}\mathbf{O}(\vec{r}_o)\mathbf{A}\left(\frac{a}{z}\vec{r}_i+\frac{b}{z}\vec{r}_o\right)\cos^3(\theta)\,d^2\vec{r}_o + t\int_{\vec{r}_o}\mathbf{O}(\vec{r}_o)\cos^3(\theta)\,d^2\vec{r}_o \tag{5.26}$$

The first term is a fraction (1-*t*) of the one discussed so far. After applying arguments similar to those presented above, we found that the second term does not change the shape of the artifacts, but intensifies them. This should be expected, because in the previous derivation the effect of our approximations was to ultimately consider **A** a constant, leading to terms in the form of the second term of eq. (5.26). This does not mean that the shape of the artifacts is independent of **A**. In fact, the dependence on the array comes back through **G**, which is intimately and uniquely related to **A**.

Background nonuniformity

The theory can also be extended to the case of background nonuniformity, which is also relevant to far-field applications. In these problems the recorded pattern is:

$$\mathbf{R}(\vec{r}_i) \propto \iint_{\vec{\xi}}\mathbf{O}'(\vec{\xi})\mathbf{A}'(\vec{r}_i-\vec{\xi})\,d^2\vec{\xi} + \mathbf{B}(\vec{r}_i) = \mathbf{O}'*\mathbf{A}'+\mathbf{B}(\vec{r}_i) \tag{5.27}$$

which, with use of eq. (5.11), upon decoding becomes:

$$\hat{\mathbf{O}} \propto (\mathbf{O}'*\mathbf{A}')\otimes\mathbf{G} + \mathbf{B}(\vec{r}_i)\otimes\mathbf{G} = \Re(\mathbf{O}') + \mathbf{B}(\vec{r}_i)\otimes\mathbf{G} \tag{5.28}$$

The artifacts are, then, given by $\mathbf{B}(\vec{r}_i)\otimes\mathbf{G}$, which is constant in the common assumption of uniform background. Otherwise **B** can be expanded in Taylor series and first and second order contributions are found again. For example, in the case of a linearly varying background:

$$\mathbf{B}(\vec{r}_i) = \vec{s}\circ\vec{r}_i + q \tag{5.29}$$

where $\vec{s}$ and q are constants, the artifact has the shape:

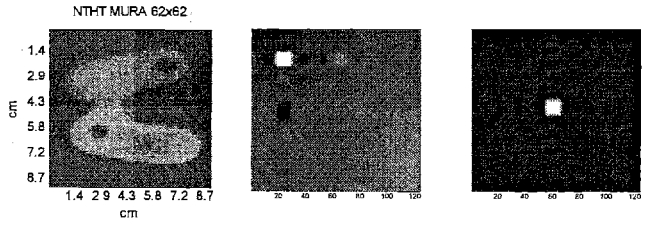
$$\hat{\mathbf{O}} = \iint_{\vec{r}_i} \mathbf{B}(\vec{r}_i)\,\mathbf{G}(\vec{r}_i + \vec{\eta})d^2\vec{r}_i = \vec{s} \circ \iint_{\vec{r}_i} \mathbf{G}(\vec{r}_i + \vec{\eta})d^2\vec{r}_i + \text{const} \tag{5.30}$$

which has the form of eq. (5.19) despite originating from a completely different term.

5.5 Verifying the near-field artifact theory

5.5.1 Simulation results

The first batch of simulations aimed at reproducing first order artifacts. Results for a NTHT MURA mask are shown in Figure 5.8. The centering artifact is evident in the simulation of the thyroid test object. The effect is very obvious despite the fact that the object is fairly well centered in the field of view, which tends to attenuate the artifact. To show the plane seen at the top right of Figure 5.5 the mask pattern must be centered and an off-centered object used. A square source close to the top left corner of the field of view was simulated. As expected, a linear modulation of the brightness appears in the image: the bottom right corner of Figure 5.8b is brighter, brightness decreasing linearly as the top left corner is approached. The slope of this plane depends on the position of the object as well as on the rotation of the mask and can be predicted from eq. (5.19). Some other structures are seen to the right and below the object. These could be due to the approximations made, in particular that about the uniformity of the object, which in this case is probably not valid. Finally, if the object is moved to the center of the field of view, the artifact, as expected, disappears (Figure 5.8c).



a) b) c)

Figure 5.8: simulation of first order artifacts. The mask was a 62×62 NTHT MURA. (a) centering artifact for a non-pattern-centered mask. For a pattern centered mask: (b) first order artifact for off-center object (c) if the object is centered the artifact vanishes.

To reproduce second order artifacts, projections of the thyroid test object were simulated for different patterns (after pattern centering). The results are in the last two columns of Figure 5.9 (for an explanation of the other columns see section 5.6). Artifacts match the shape predicted by theory and shown in Figure 5.7.

5.5.2 Published results

Results of earlier work on coded apertures appear in literature. In Figure 5.10 are shown two coded aperture images of a hand and of an ECT cold rod phantom from ref. [18]. We know that the authors have used a 1-mm-thick lead mask with 1.5-mm holes arranged in a URA pattern, but we do not know the dimensions. The artifacts for this family are almost identical to those expected from NTHT MURAs (see Figure 5.6 and Figure 5.7).

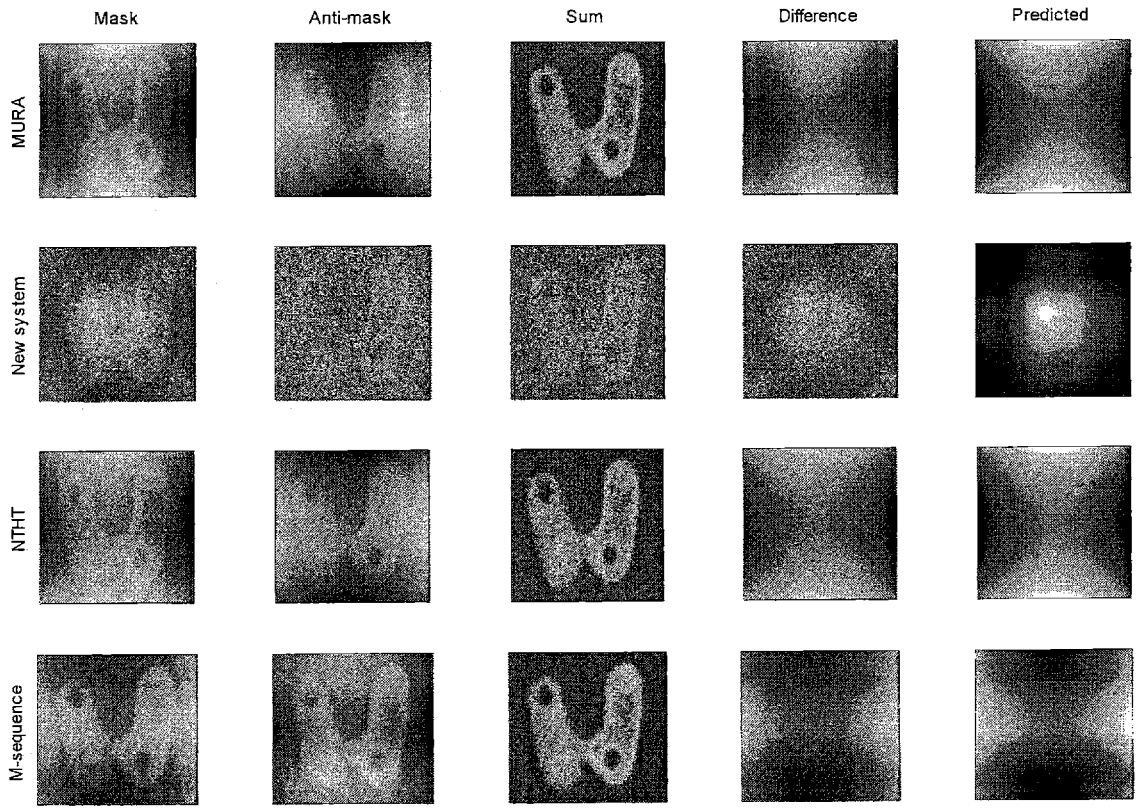


Figure 5.9: simulation results. Four arrays were considered: a 79 × 79 MURA, a 77 × 77 NS array (a product array similar to the 33×33 PNP of Figure 5.6), a 62 × 62 NTHT array based on a 31 × 31 MURA and a 63 × 65 *m*-sequence. The five columns show, for each mask, the two views, the sum and the difference picture (see section 5.6), and the prediction of the artifact according to eq. (5.25). Note the poor SNR for the NS mask and, especially, anti-mask, as predicted in §4.4.2.

We believe the large background of the hand image to be the plane predicted by theory and seen in our simulations. Such artifact is replaced by second order artifacts in the phantom images, consistently with our observations on the centering of the object. In this image, a horizontal line is also evident. Unfortunately, we do not know the shift of the pattern used in that study, or whether it was different from that used for the hand image. Here we can only note that it would be consistent with a first-order-centering artifact due to mask centering along one dimension only. The presence at the same time of a first and a second order artifact could be an indication that the center of mass of the object was close to, but not exactly on, the center of the field of view. In this situation first order artifacts would be reduced to the point that second order artifacts would become visible.

The artifacts seen in the images of Figure 5.10 are consistent with the predictions of the theory.

5.6 Near-field artifact reduction

We have already pointed out that, given a pair (**A**, **G**) with ideal correlation properties, also its negative, i.e. the pair (**1-A**, **-G**) has ideal correlation properties. In **A** closed elements are substituted with open elements and vice versa (section 2.4). This is also called the anti-mask (e.g. [55]). When the image is reconstructed, the two sign changes cancel in eq. (5.11), so that the reconstructed object does not change sign, i.e. we are not taking a "negative" picture:

$$\hat{\mathbf{O}} = [\mathbf{O}' * (\mathbf{1} - \mathbf{A}')] \otimes (-\mathbf{G}) = \Re\{\mathbf{O}' * [(-\mathbf{G}) \otimes (\mathbf{1} - \mathbf{A}')]\} = \Re[\mathbf{O}' * (\text{const} + \delta)] = \Re(\mathbf{O}') + \text{const} \quad (5.31)$$

On the other hand, first and second order artifacts are seen to depend only on **G** (eq. (5.19) and (5.25)), and must change sign. So, when an image of the same object is taken with an anti-mask, artifacts change sign while the reconstructed object does not. Adding the two images should cancel artifacts and reinforce the reconstruction, while subtraction of the two pictures should cancel the object and reinforce artifacts. We simulated this experiment with the test object: the remarkably successful outcome is shown in Figure 5.11.

In Figure 5.9 the same experiment was repeated with equal success for different array families. From these simulations we learn that, when applying this artifact reduction technique, one should be aware of the noise properties of both mask and anti-mask in the choice of the mask pattern. For instance, even if low-throughput arrays such as the NS arrays of §2.4.5 have poor but still reasonable noise properties, the associated anti-mask is very noisy (section 4.4), causing a considerable signal-to-noise loss in the reconstructed image. In our experience, (M)URA or (M)URA-based patterns provide the most

Figure 5.10: a hand (a) and a ECT cold rod phantom (b). In this phantom the smallest rod diameter is 6.4 mm. From ref. [18].

balanced performance. Note that an NTHT (M)URA can be considered a (M)URA with partially closed holes. Its inverse, therefore, is a (M)URA with partially closed holes. The result is the same as inverting only the original positions of the NTHT (M)URA. Blank lines, therefore, are not inverted, which explains why the negative of an NTHT (M)URA can be made of the same density of its 12.5% open positive. From a point of view of the decoding array, the change of sign of blank lines is irrelevant, because they correspond to 0s of **G** (§2.4.7).

5.7 Experimental results

To verify our simulations, a series of experiments were carried out using a thyroid phantom in conjunction with a Siemens E-Cam. The results from the exposure of a 2d phantom injected with ~200 μCi of ^{99m}Tc are reported in Figure 5.12. With the phantom about 40 cm from the detector and approximately 9 cm from the mask, distances suggested as optimal by our simulations (see Chapter 6), the two exposures took about 8 minutes each. Agreement with the computer simulation of Figure 5.11 is satisfactory.

5.8 Mask thickness artifacts

A thick mask stops rays that ideally pass. This generates artifacts because the pattern projected on the detector is altered and can not be decoded properly. These artifacts are difficult to predict theoretically and were investigated by computer simulation. The result is in Figure 5.13. Artifacts are not unbearably disruptive, but do generate false peaks in the sidelobes, in vertical and horizontal alignment with the source. For a constant mask thickness, smaller pinholes generate worse artifacts because the acceptance angle of the mask is even smaller. The optimal mask thickness must be chosen with a compromise with

the SNR. In fact, if a mask is made too thin to avoid artifacts, statistical noise may be such that false peaks from thickness artifacts will always be lost in noise and there would be little point in not increasing the mask thickness. A way of reducing this effect would be to move the detector away from the source, but this would attenuate artifacts at the expense of collection efficiency and, thus, of the SNR.

The problem of finding the optimal mask thickness is treated at length in Chapter 6 and Chapter 7, where a quantitative approach to solving the tradeoff between noise and artifacts is developed.

5.9 Summary

The arrays that provide an ideal point spread function in far-field coded aperture imaging do not operate under ideal conditions in near-field imaging. The images produced are affected by artifacts that have already been encountered by other researchers and are found in published results. A theory capable of predicting the shape of such artifacts has been developed, providing valuable insights.

A few image improvement strategies are suggested. Centering the object and the mask pattern were shown to eliminate first order artifacts, the strongest, but are not effective in eliminating residual

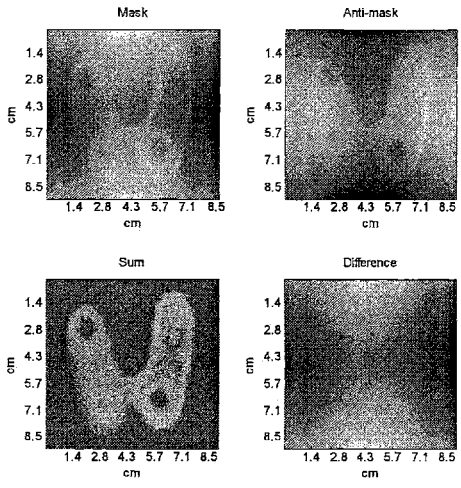


Figure 5.11 simulation of two exposures of a thyroid: mask (top left) and anti-mask (top right). Note that artifacts change sign. When the two images are added (bottom left) they cancel out and the signal is reinforced. If the images are subtracted, the opposite is true and the artifact is reinforced, while the object cancels out (bottom right).

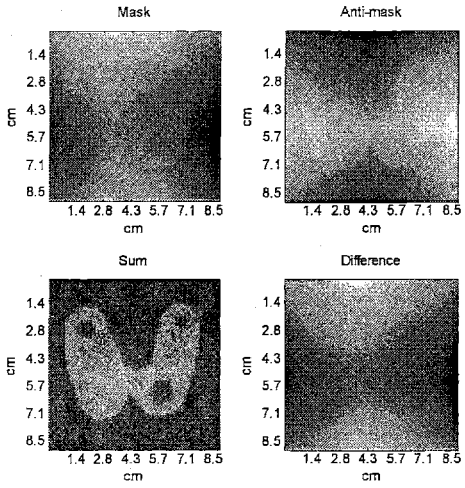


Figure 5.12 experimental results for a thyroid phantom. Figure 5.11 provides comparison with simulation. To obtain a truly 2d image only the bottom of the phantom was filled. The spike coming out of the bottom left lobe is the injection channel. The measured resolution of this image is about 1.5 mm.

second order artifacts. The use of a technique previously suggested ([55]-[58]) for non-uniform background reduction proved successful: two images are taken, one with a mask and one with its anti-mask, and then added. The exposure time, divided in two halves, is not increased by more than the time necessary to physically change the mask pattern and start a new acquisition. Furthermore, if the mask is anti-symmetric (a pattern in which some rotation or reflection results in the replacement of open with closed mask positions and vice versa), the anti-mask is simply a rotation or reflection of the mask and there is no need to fabricate two masks. Also, changing mechanisms, when needed, are greatly simplified ([55]).

The price paid is a reduced range of available patterns: in our case we were forced to choose a 62×62 pattern in place of a 74×74, with a 17% loss in resolution at constant field of view (see section

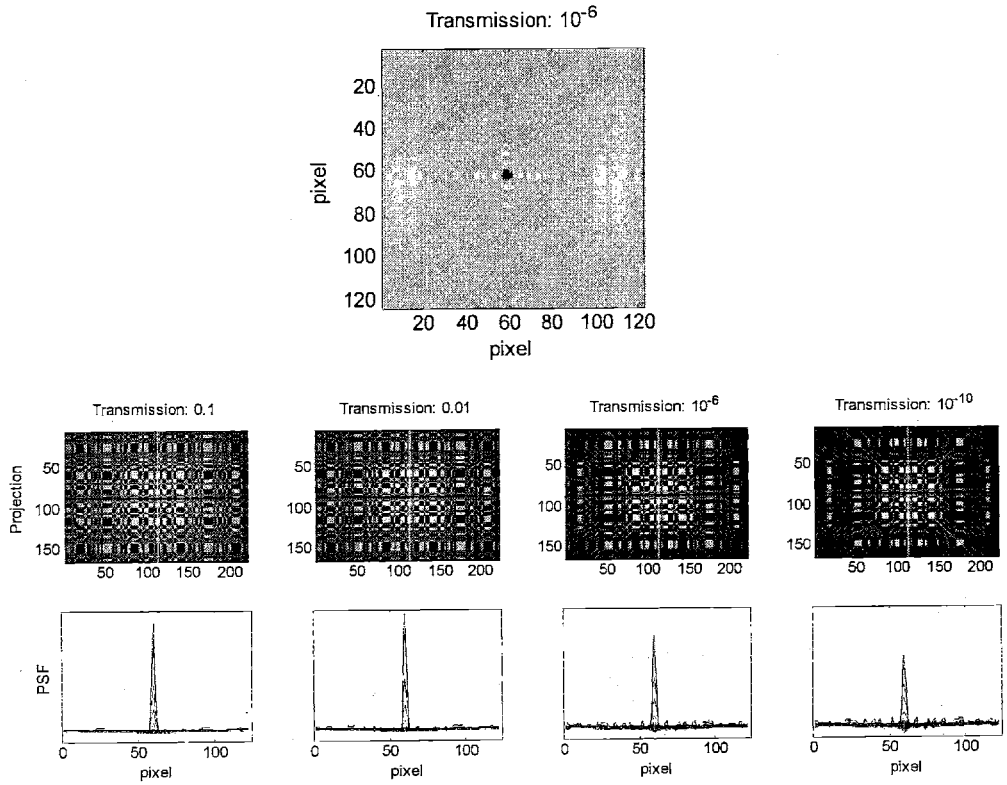


Figure 5.13: simulation results. Top: PSF of a thick mask. Contrast greatly enhanced to show artifacts. A quantitative idea can be obtained by looking at the plot of the same PSF in the bottom row. Bottom: Mask thickness affects the shape of the projection of the mask onto the detector. Small holes at an angle from the center disappear from the projection when mask thickness is increased. The PSF deviates from ideality. Mask used MURA 61×61 with $p_m = 1.128$ mm. Very high activity, $a = 9.41$ cm, $z = 40$ cm. Mask made of tungsten. Isotope: ^{99m}Tc. Thickness: 0.763, 1.527, 4.580 and 7.633 mm. Zero order correction applied. The central 122×122 portion of the projection shown above was used for decoding.

6.1).

A factor expected to strengthen artifacts is the field of view. In this application we were mainly concerned with high resolution. For a detector of given dimensions, this results in a relatively small field of view (see section 2.6). In applications with a wider field of view, expansion to the second order may not be sufficient, but one can also argue that the method can be extended to higher orders, and the ensuing artifacts may turn out to be eliminated with the same mask / anti-mask technique. This method has indeed already given a proof of robustness. In fact, collimation effects from finite mask thickness, even if included in our simulations, were not taken into account in the theory. However, both simulations and experiments show that taking two pictures seems to overcome this additional difficulty.

Chapter 6 DESIGN AND FABRICATION OF A CODED APERTURE: EXPERIMENTAL RESULTS

In this chapter all ideas, concepts and methods of the previous chapters are implemented in the design of the coded aperture for a high-resolution camera. The aim is to provide not only a set of technical specifications but also a rational design procedure, that determines mask parameters one at a time from sound principles.

Experimental results follow. Experimental practice showed that it is very tedious to have to achieve and maintain perfect focusing. In section 6.3 is described an empirical technique to decode projections for non-integer values of the sampling parameter α, which makes the issue of correct focusing, if not irrelevant, certainly less critical.

6.1 Mask design

The mask was designed to optimize resolution over a 9 × 9 cm field of view. For the design of the first coded aperture, not wishing to push any design limits, the E-Cam was set up to make 222 × 161 pixel pictures with a pixel size $p_d = 2.398$ mm. With these choices, enough data are available to determine the mask pattern and several dimensions.

6.1.1 Material and technology

From the discussion of mask thickness artifacts (section 5.8), the best material to fabricate the mask is the one having minimum thickness for a given attenuation. From the theory of γ-ray attenuation, this is the material with the maximum product $\mu \times \rho$, with μ attenuation coefficient and ρ the density of the material. The best material is uranium (48.97 cm^{-1} at 140 keV), followed by platinum (38.4 cm^{-1}), gold (35.9 cm^{-1}), tungsten (30.5 cm^{-1}) and lead (22.96 cm^{-1}). The choice is dictated by practicality, availability, cost and fabrication technology. We picked tungsten and, since it is very hard to machine, we chose to photo-etch the pattern.

6.1.2 Pattern family

The drawback of photo-etching is that non-self-supporting arrays are very difficult to make (if at all) because the mask is built from very thin layers (which will be stacked later), each of which has the shape of the mask. This ruled out the use of the array family with the highest SNR, the (M)URA. *m*-sequences can be folded in self-supporting 2d arrays (§2.4.3), but, from artifact theory, we would like to have an anti-symmetric mask. NTHT (M)URA patterns with $e = 2$ (see §4.3.3) are the family with the next-highest SNR. Candidate patterns still have to be anti-symmetric: when used in an NTHT pattern URAs and MURAs still offer the same SNR. URAs have axial symmetry: after rotation about the symmetry axis, its points do not change place and, thus, do not change sign. For MURAs, which are symmetric about a center, the problem is limited to a point, the center of the rotation. Consequently we chose a NTHT MURA mask pattern. The pattern is going to be shifted so that the center of symmetry is at the center of rotation. This means that the totally closed and open line at the center of the original MURA must be at the center of the pattern.

6.1.3 Geometric parameters

From sampling considerations (see §5.2.1) α was set to at least 2, i.e. the projection of each mask hole must cover 2 detector pixels or more. Maximum resolution for a given field of view is obtained for the highest number of mask pixels (eq. (2.91)). Since at most 161 pixels of the detector can be used, using the minimum α is equivalent to choosing the maximum number of mask elements, which is, since $\alpha = 2$, $161 / 2 = 80.5$, i.e. 80. Since NTHT MURAs with maximum SNR have $e = 2$, the MURA on which it is based must be at most 40×40. The side of a MURA must be a prime number. The largest prime number smaller than 40 is 37, but the 37×37 MURA is symmetric, not anti-symmetric. The next array is a 31×31 and happens to be anti-symmetric. This was our final choice. The corresponding NTHT is a 62×62 array ($N_T = 3844$, $N = 480$, see §2.4.4) projecting on the 222×161 pixels available. Not all of the active area of the detector is used, a limitation coming from the decision to keep α integer to avoid sampling artifacts (see §5.2.2). The effective area used is $d_{d,\,eff} = 62 \times 0.2398$ cm = 29.74 cm. Data outside this area will be discarded, so this is the value to be substituted in eq. (2.79) to get m:

$$m = \frac{29.74}{9} + 1 = 4.304 \tag{6.1}$$

The mask pixel size is then determined from eq. (2.96):

$$p_m = \frac{\alpha d_p}{m} = 1.114 \text{ mm} \tag{6.2}$$

The geometric resolution is given by eq. (2.87):

$$\lambda_g = \frac{m\, p_m}{m-1} = 1.45 \text{ mm} \tag{6.3}$$

to be compared against the 3.7-mm intrinsic PSF limiting the resolution of the Anger camera. This value must then be combined with collimator parameters to obtain system resolution. Specifications of the collimators equipping the E-Cam indicate that the high-sensitivity collimator has 2.405 cm-long hexagonal holes with a side-to-side distance of 2.54 mm. For ^{99m}Tc, it has 14.6 mm geometric resolution at 10 cm from the detector, while the ultra-high-resolution collimator (3.58 cm-long hexagonal holes with a side-to-side distance of 1.16 mm) has 4.6-mm resolution at 10 cm. The price is paid in terms of sensitivity: 1063 vs. 100 cpm / μCi. A 4-mm pinhole, the smallest available, has a reported geometric resolution of 6.2 mm and a sensitivity of 123 cpm / μCi.

To estimate the sensitivity of the coded aperture we first need to determine the object-to-detector distance. In fact, the design process has so far fixed only the magnification coefficient, which determines the ratio of the mask-to-object and mask-to-detector distance (respectively, a and b) but not their sum z. In the determination of this variable, many tradeoffs are involved. Low z leads to higher count rates and, thus, better statistics, but mask thickness artifacts are enhanced (section 5.8). On the other hand, if the mask is too thin, the SNR of the image, which depends on the particular object at hand, decreases. The issue is very involved: an empirical approached was chosen to determine mask thickness and object-to-detector distance at the same time.

6.1.4 Determination of mask thickness and object-to-detector distance

The ultimate factor in determining optimal performance is the quality of the image. This is affected by random noise, which can be quantified in terms of the SNR, and the presence of artifacts. Building on the observation that thickness artifacts look in many ways like random noise, it seemed reasonable to forget the distinction of the two and define a figure of merit (FoM) as the normalized second moment of inertia of the peak reconstructed from a point source. Mask thickness causes the peak to broaden and others to appear, while counting statistics causes noise peaks in the sidelobes to increase: in both cases the moment of inertia of the central peak increases. The mathematical definition of the FoM is:

$$FoM = \sqrt{\frac{\iint \left[(x-\mu_x)^2 + (y-\mu_y)^2\right] \hat{\mathbf{O}}(x,y)\,dx\,dy}{\iint \hat{\mathbf{O}}(x,y)\,dx\,dy}} \tag{6.4}$$

with $$\mu_x = \frac{\iint x\,\hat{\mathbf{O}}(x,y)\,dx\,dy}{\iint \hat{\mathbf{O}}(x,y)\,dxdy} \quad \text{and} \quad \mu_y = \frac{\iint y\,\hat{\mathbf{O}}(x,y)\,dx\,dy}{\iint \hat{\mathbf{O}}(x,y)\,dxdy} \tag{6.5}$$

where $\hat{\mathbf{O}}(x,y)$ is the brightness of the reconstructed image of a point source. Note that this definition implies that the lower the FoM, the better the result.

Before using the FoM, its efficaciousness was tested in a series of trials targeting its dependence on different parameters. The mask under consideration has the characteristics described in the previous section. In Figure 6.1 is shown the dependence of the FoM on the object-to-detector distance. The mask thickness was kept constant and the object activity set to a very high level to avoid interference with Poisson noise. The FoM decreases, i.e. improves, with distance. The effect on the reconstruction is also shown in three sample reconstructions. Of course, artifacts have some structure, but are fairly well

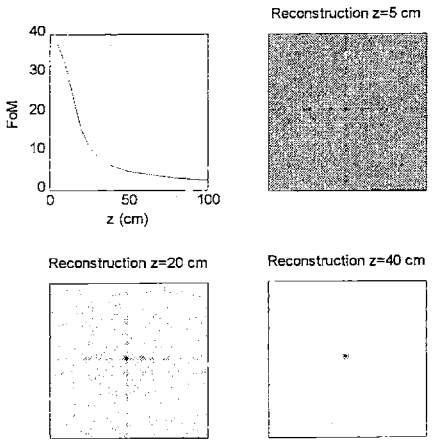


Figure 6.1: NTHT MURA 62×62 mask. Figure of Merit as a function of z at constant transmission ($t = 0.01$, mask thickness = 1.524 mm) and infinitely good statistics. As expected, at low z image quality is degraded due to thickness artifacts. At $z = 40$ cm little artifacts are seen. Other simulation parameters: $\alpha = 2$, $n = 62$, $p_m = 1.114$ mm, detector effective area 29.74×29.74 cm, $p_d = 2.398$ mm.

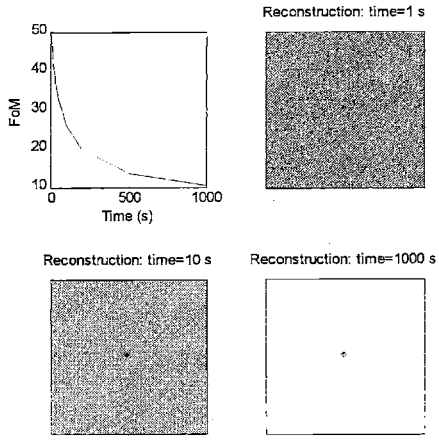


Figure 6.2: Figure of Merit as a function of exposure time at constant transmission ($t = 0.01$, mask thickness = 1.524 mm), object-to-detector distance ($z = 40$ cm) and source activity (1 μCi). As time increases the source stands out more sharply on background noise and the FoM decreases. Other simulation parameters as in Figure 6.1.

distributed over the image and are very similar to random noise in raising the sidelobe value relative to the peak. In Figure 6.2 mask thickness and object-to-detector distance are fixed and exposure time is varied. As expected, the FoM decreases for increasing exposure time.

To determine the best mask thickness and object-to-detector distance a set of simulations calculated the FoM for a wide range of thickness and distance. The result is in Figure 6.3. The absolute minimum is obtained for a distance of 30 cm and a mask thickness of about 1.2 mm. We decided to adopt $z = 40$ cm because the curve has a wider plateau, which should be more forgiving of the approximations of the method. A mask thickness of 1.52 mm was chosen, corresponding to 1% transmission for ^{99m}Tc. This higher value seemed more attractive also because the mask would be thick enough to be useful for higher-energy isotopes as well. For comparison, the septal penetration of collimators is also in the order of 1%, depending on resolution. Note that for increasing distance, the FoM minimum occurs at higher thickness. This is understandable, because a higher z is more tolerant of mask thickness and so favors better opacity to increase the SNR diminished by the higher distance.

Once z is determined, a and b are easily found:

$$\begin{cases} a+b=z \\ m=1+\dfrac{b}{a} \end{cases} \Rightarrow \begin{cases} a=\dfrac{z}{m} \\ b=z\dfrac{m-1}{m} \end{cases} \tag{6.6}$$

A lower limit on z may be imposed through a by the need to allow some space between mask and object plane to allow placement of the sample. In this case, there is little worry because $a = 9.29$ cm and

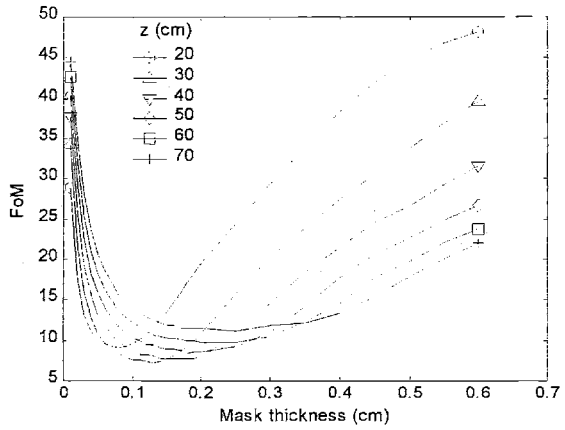


Figure 6.3: Figure of Merit as a function of mask thickness for different object-to-detector distances. Activity 10 μCi, exposure time 300 s. As expected, the FoM has a minimum. Other simulation parameters as in Figure 6.1.

$b = 30.71$ cm.

The final characteristics of the mask are summarized in Table 4.1, where they are compared to those of a mask previously built by this research group.

6.1.5 Calculation of mask sensitivity

With all dimensions determined it is possible to estimate the sensitivity of the coded aperture by adding the flux through all pinholes. The solid angle subtended by a pinhole on the instrument axis is given by ([59]):

$$\Omega = a\int_{-p_m/2}^{p_m/2}\int_{-p_m/2}^{p_m/2}\frac{dx\,dy}{\left(x^2+y^2+a^2\right)^{3/2}} = 4\arctan\left(\frac{\frac{p_m}{2}\frac{p_m}{2}}{a\sqrt{\left(\frac{p_m}{2}\right)^2+\left(\frac{p_m}{2}\right)^2+a^2}}\right) \tag{6.7}$$

Substituting the values: $\Omega = 1.44 \times 10^{-4}$. Assuming this constant for 480 pinholes, the fraction of emitted photons passing through the mask is $480\ \Omega / 4\pi = 5.49 \times 10^{-3}$. An activity of 1 μCi $=3.7 \times 10^4$ Bq is equivalent to 2.22×10^6 cpm, so the sensitivity is about 12200 cpm / μCi.

Mask pattern	NTHT MURA 62 × 62	URA 43 × 41
Open fraction	12.5%	50%
Mosaicked	yes	no
Self-supporting	yes	no (aluminum back pane used)
Mask pixel size	1.11 mm	4 mm
Resolution (at FoV = 9 cm)	1.5 mm	9 mm
Mask symmetry	anti-symmetric about center	even (horizontally) / odd (vertically)
Material	Tungsten	Lead
Fabrication technology	Photo-etching	Computer controlled milling
Thickness	1.5 mm	3.17 mm
Attenuation at 140 keV	99%	99.3%

Table 6.1: summary of the properties of the prototype mask. Characteristics are compared to those of another mask previously built by this research group.

This is a very rough estimate because it does not take into account mask transmission, thickness and the fact that Ω is not the same for all pinholes. The first effect is the only one that can be calculated. Since the seven eighths of the mask that are supposedly opaque are in reality 1% transparent, the number of counts is 7% higher (1% for each of the 7 opaque positions corresponding to an open position). With this correction the expected counts are about 13000 cpm / μCi. The other two corrections are more difficult to calculate and were evaluated with the help of the computer code. They both go, however, in the sense of a reduction of the expected number of counts which, in fact, turns out to be about 8150 cpm / μCi. Sensitivity can then be used in the expression of the SNR to determine, with object and background parameters, the SNR (see section 4.7).

6.2 Experimental results

The first round of experiments was carried out on July 18, 2000 on a Siemens E-Cam at the Dana

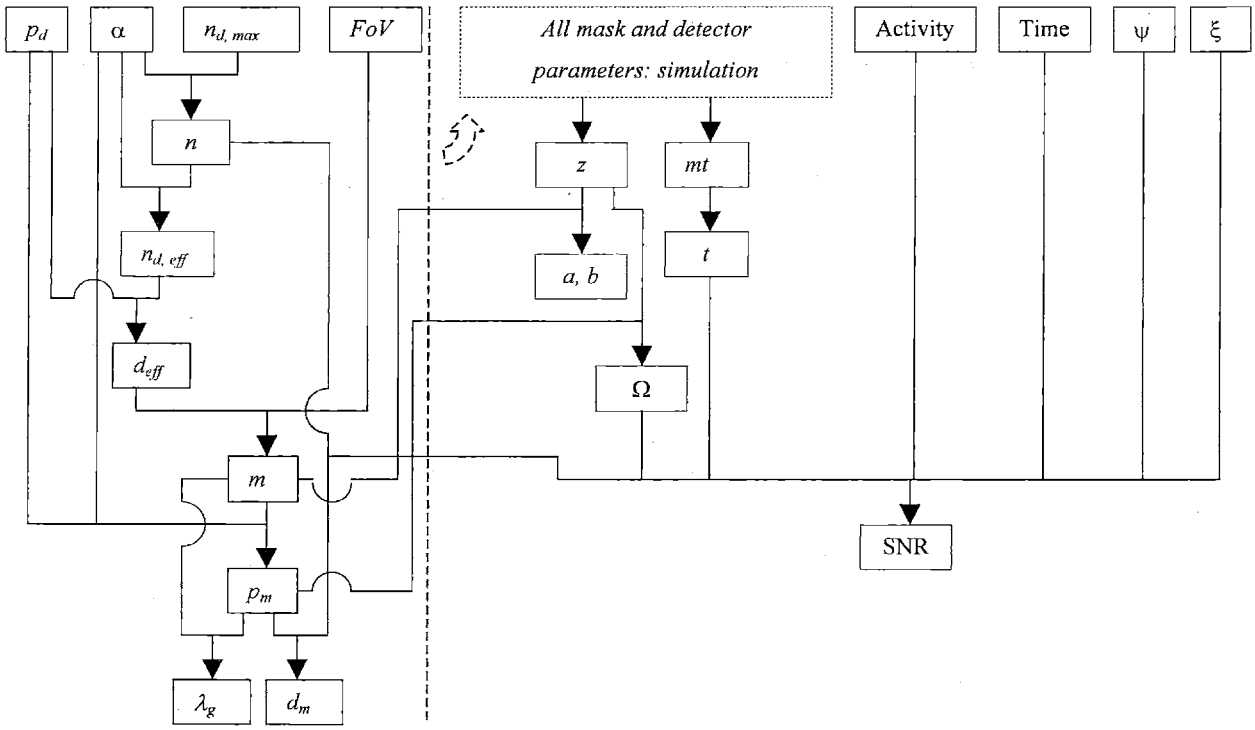


Figure 6.4: flow chart showing the procedure to determine all the design parameters. On top are the parameters imposed or chosen but not constrained by others. $n_{d\,max}$ is the maximum number of detector pixels (in this case 161). *mt* is the mask thickness. All parameters to the left of the dashed line feed into the simulation and affect, at least potentially, the determination of *z* and *mt* (dashed arrow).

Farber cancer institute in Boston. The first objective was to measure the PSF of the mask built. In all experiments ^{99m}Tc was used.

Unfortunately it was not possible, due to a mechanical constraint, to locate the mask 40 cm from the detector. The maximum z achievable was 38.6 cm. b was recalculated to be 29.62 cm and a 8.98 cm. Even if the difference is small, it is important to keep track of it because α is very sensitive to variations in these parameters (section 2.8) and the reconstruction very sensitive to variations in α (§5.2.2). Due to measurement imprecision, the first decoded image revealed that $\alpha = 2.02$ and the PSF was not as good as it could have been (Figure 6.5). The procedure presented in section 2.8 was used to fine tune the instrument until the result of Figure 6.6 was reached.

With the instrument focused, the next objective was to test resolution. Two capillaries, 7.5-cm long, with an internal diameter of 1.15 ± 0.05 mm were filled (except some distance at the extremities to insert cotton plugs against spills) with 150 μCi each. Two pictures were taken, each for a total of 2 million counts. Each exposures took 80 seconds. The result is in Figure 6.7.

Artifact reduction works as predicted. The strips in the background running parallel to the sources are due to noise correlations and the nature of the object (see section 4.5). Interestingly enough the effect disappears in the difference image, where other artifacts are evident. Resolution can be appreciated in the very sharp line representing the capillaries. The closest point between the capillaries was a 3-4-mm gap

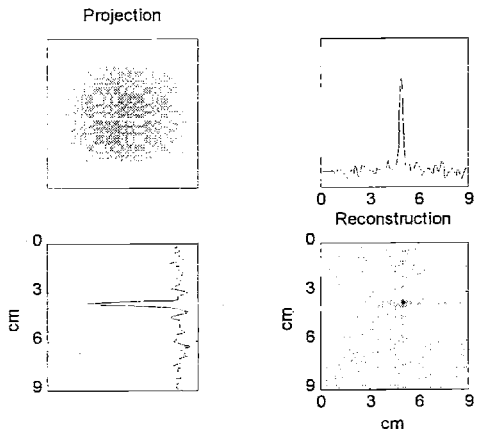


Figure 6.5: experimental PSF. 800k counts in 44 s. Projection, reconstruction and cross sections of the reconstruction through the peak are shown. The projection showed that $\alpha = 2.02$. Some fine tuning was necessary.

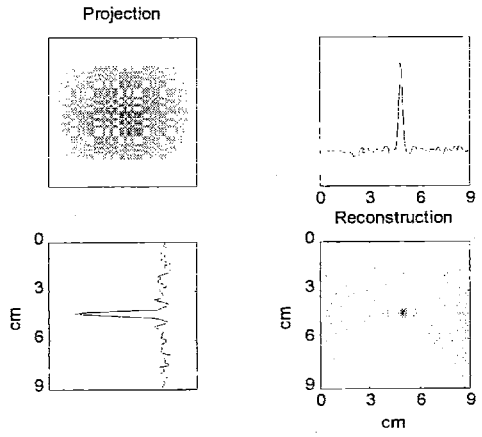


Figure 6.6: experimental PSF. 1.6M counts in 736 s. The projection showed that $\alpha = 2$. The PSF improved sensibly. Meanwhile, during adjustment, the center of the field of view was aligned the mask rotated. The technologist was also able to prepare a smaller point source. Of course, none of these changes affects focusing.

and is clearly resolved. Also note the aliasing of the top of the left capillary, that exits the picture from the top left corner to reappear to the top right. More care should have been taken in blocking the PCFV or positioning the source within the FCFV. However, the result is not a disaster, proving the robustness of the technique. To get a better measurement of resolution three capillaries were used. They were taped side by side and only the outer ones were filled. The center tube was used as a spacer. Including a wall thickness of 0.2 ± 0.02 mm, the distance between the edges of the active regions was 1.95 ± 0.13 mm while the center-to-center distance was 2.9 ± 0.08 mm. It is evident that resolution is better than both these limits.

The object of the next experiment was a thyroid phantom, filled only at its bottom to avoid object thickness effects. The phantom was placed so that the activity was equal everywhere, except at the cold spots that had no activity. For this object the maximum ψ_{ij} is 5.6×10^{-4} and the average 2.6×10^{-4}. Background was also mesured and resulted in $\xi = 5.56 \times 10^{-4}$. The result is Figure 6.9. Near-field artifact reduction is still very successful and resolution is still high. All cold spots are clearly identified and a hint of the injection channel is visible at the top right, exiting from the lobe of the thyroid. As expected, noise correlations do not result in particular structures because the shape of the object is not as particular as that

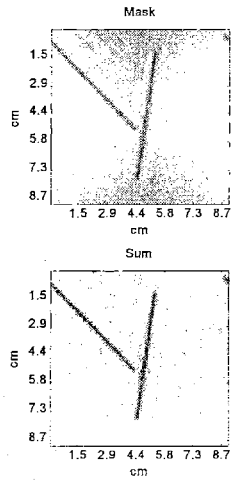


Figure 6.7: reconstruction of two capillaries. Top left: picture obtained with the mask. Top right: picture obtained with the anti-mask (rotation of the mask). Bottom left: sum image. This is the final image. Bottom right: difference image: this image shows near-field artifacts. It is important to verify that the object disappears, which indicates correct alignment of the images of the top row.

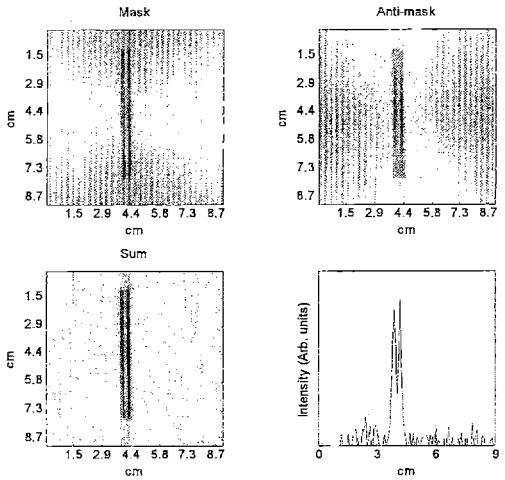


Figure 6.8: reconstruction of two parallel capillaries. At the bottom left is a plot of the brightness along the horizontal section at 4.4 cm. Two peaks are clearly resolved. They are separated by 4 pixels, i.e. 4×2.398 mm at the detector. Dividing by m-1 one can get the object space equivalent, which is 2.9 mm. See text for capillary specifications. Two exposures of about 80 s for a total of 2 million counts each.

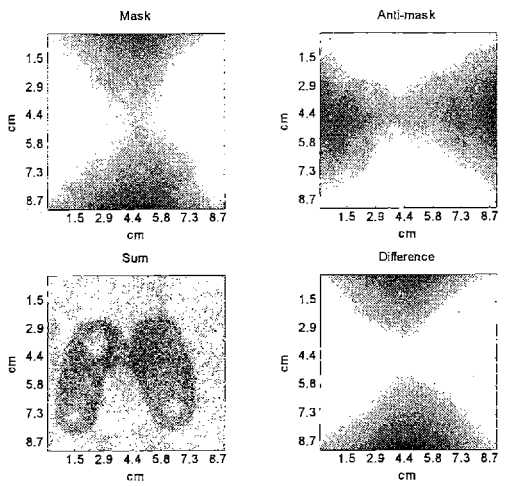


Figure 6.9: thyroid phantom. Only the bottom was filled to have a 2d phantom. Two reconstructions. 350 μCi injected. Two exposures, each for 20 million counts (about 500 s).

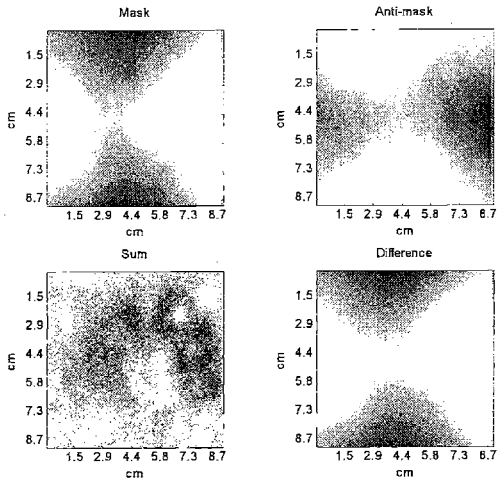


Figure 6.10: the phantom was filled completely to see the effects of out-of-focus planes on the image. Due to a focusing error the reconstruction is severely blurred. 20 million counts for each projection.

of a line source. These results can be compared to those of references [29], [40] and [42] where cold spots are barely recognizable.

The last experiment of the day was to fill the phantom to see the consequences of imaging a three-dimensional object which would include out-of-focus planes. Before taking the image the system was refocused because filling up the thyroid causes its average plane to move away from the detector. Unfortunately, as can be seen in Figure 6.10, the compensation was wrong.

6.3 Electronic focusing

The focusing mistake was fixed by simply resampling the data with linear interpolation[14]. In other words, it is assumed that the real data, as a continuous function of space, can be recovered from the values stored in the pixels by assuming a linear distribution of brightness between pixels. This continuous function can then be sampled with any pitch, giving an estimate of what the pixel counts would have been if pixels had been on a grid having this new pitch. The approach is crude, but we found it to work rather well. The implication is that decoding can be carried out for any value of α, not only for integer values.

[14] Other resampling methods (bicubic and spline) were tried but did not seem to provide any advantage.

In fact, resampling the data with a different pitch means assuming that the detector pixel dimension is different. With this trick, we can restore α to an integer value, α', that can be set to the integer closest to α. The resampling interval is obtained setting (see eq. (2.96)):

$$\alpha' p'_d = \alpha p_d = m\, p_m \tag{6.8}$$

which gives:

$$p'_d = \frac{\alpha\, p_d}{\alpha'} \tag{6.9}$$

where α is measured directly on the acquired image. Of course, the fact that α is not what we expected, means that m must be different from its design value, because p_m and p_d are fixed. The real (experimental) value of m can be recalculated by using:

$$m = \frac{\alpha\, p_d}{p_m} \tag{6.10}$$

The deviation from design conditions can be due to a, b or both (eq. (2.74)). In our case this is of little interest, but there is a case in which this information is fundamental. If the object-to-detector distance is not known, the α needed for the correct reconstruction is not known. One can then decode trying a range of different α and choose the sharpest reconstruction as the good image. Possible values of α range from p_m / p_d (because m is at least 1, so the minimum size of the projection of a hole is the size of the hole itself) to the ratio of the number of active detector pixels to the number of mask pixels n. From knowledge of the successful α one can find an estimate of m and, from it, of a. From (eq. (2.74)):

$$a = \frac{b}{m-1} = \frac{b}{\frac{\alpha' p'_d}{p_m} - 1} = \frac{b}{\frac{\alpha p_d}{p_m} - 1} \tag{6.11}$$

because only when the sampling grid $\alpha' p'_d$ matches the pitch of the mask projection $m\, p_m$ can the reconstruction be sharp.

The technique was applied to the thyroid data of the previous section. The result is in Figure 6.11. A first, coarse, scan helps focusing on the correct region which is then explored in more detail. The thyroid clearly comes in focus as α changes. The best result is obtained for $\alpha \cong 2.27$. In this image all four spots are clearly resolved. The top left spot is a hot spot, the other three are cold spots. The hot spot has

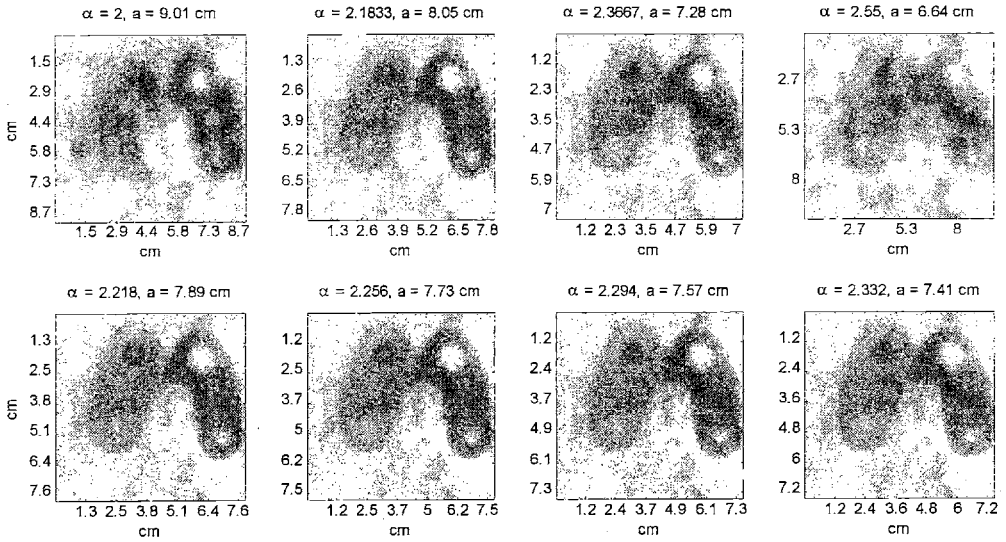


Figure 6.11: electronic refocusing. Top row: coarse scan. The focal plane seems to be between $\alpha = 2.18$ and $\alpha = 2.37$. Bottom row: four other reconstructions were decoded for α spaced so that there would be 6 reconstructions at equal intervals between 2.18 and 2.37. The focal plane seems to be somewhere around $a = 7.6$ cm.

about the same activity of the right lobe and stands out of the left lobe, which is actually a colder lobe, having 50% of the activity of the right lobe.

The result is also in good agreement with eq. (2.106). In fact, susbtituting the mask and detector parameters, with $b = 29.8$ cm and $a = 8.8$ cm (our best estimates):

$$\frac{d\alpha}{da} = -\frac{d_m b}{d_p a^2} = -.17\frac{1}{\text{cm}} \tag{2.106}$$

For the thyroid, which is 2.2-cm-thick, $d\alpha = -0.17 \times 1.1 = 0.19$, which is exactly in the order of the observed defocusing. This is an experimental proof of the sensisitivity to depth of the method. Fortunately, electronic refocusing can recover data taken with a defocused setup.

6.4 In vivo experiments

All in vivo experiments were performed at the Center for Molecular Imaging of the Massachusetts General Hospital in Boston.

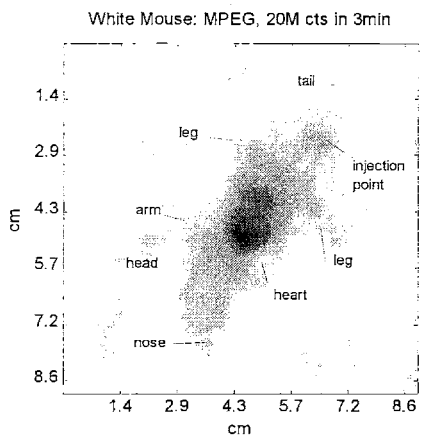


Figure 6.12: body scan of a white mouse. Blood pool agent. Total exposure time 6 min 30s. Total number of counts 40×10^6. Best estimate of the injected activity 1.5 mCi.

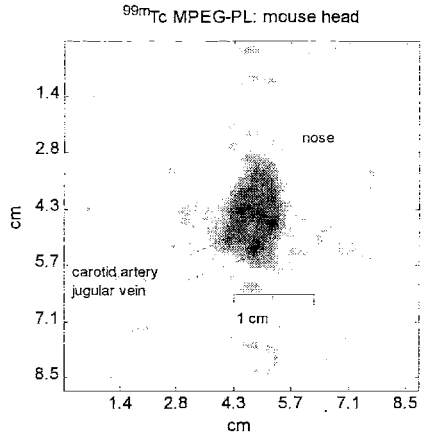


Figure 6.13: same mouse, the body below the neck was shielded with a lead sheet. Details of the head are now clearly visible. 277+279s exposure for 5×10^6 counts each.

6.4.1 Blood pool agent

The first trial was with a blood pool agent, MPEG-PL, labeled with ^{99m}Tc. First, a picture of the whole mouse was taken (Figure 6.12). The shape of the body is clearly recognizable. Most of the activity is concentrated in the trunk. Activity outside the body is also visible, but can be easily discriminated by resetting the lower threshold of the color scale. Resolution is high as compared to collimator scans, but the SNR is poor at dim points. This is expected from SNR theory, where advantages were proven to be present for bright points only.

A much more interesting picture is obtained when the body is shielded and the head is placed at the center of the field of view (Figure 6.13). The image is much brighter and the SNR improvement allows localization of details, including major vessels and vessels to and from the nose, that were invisible in the whole bodu scan. Isolation of the head, has tremendously improved ψ_{ij} for points of the head because, shielding the activity in the trunk reduced I_T. Note that this does not necessarily diminish the interest of the study, which may be focused on an area of the body only, in this case the head.

6.4.2 Bone agent

A bone agent, ^{99m}Tc-labeled MDP, was tried next. The experiment of a whole body scan (Figure 6.14) followed by a focused image was repeated. The body was shielded from the waist up to shield

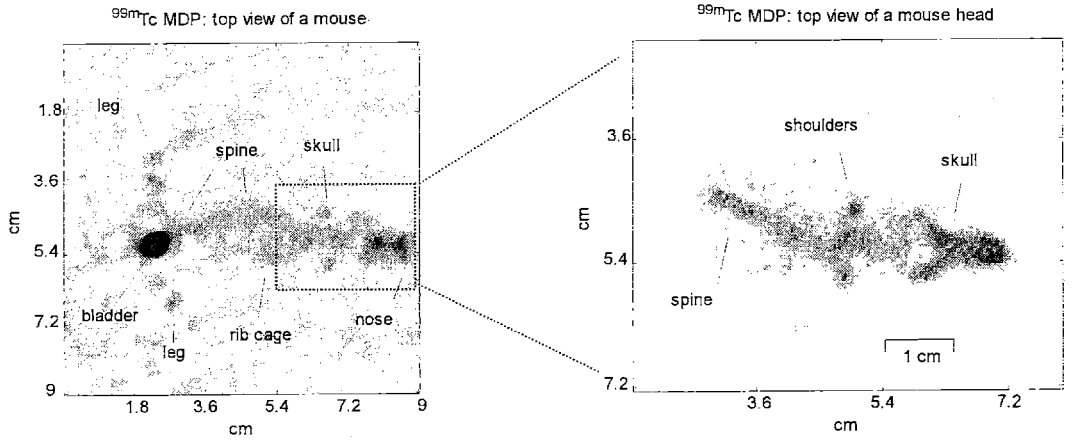


Figure 6.14: whole body scan of ^{99m}Tc-MDP in a mouse. Top view. $2 \times 4 \times 10^6$ counts collected in 490+495s. Injected activity: 100 μCi. Most of it ends up in the bladder.

Figure 6.15: after shielding the mouse from the waist down, the SNR is improved. Same mouse as in Figure 6.14. $2 \times 2 \times 10^6$ counts in 739+754s.

activity in the bladder (Figure 6.15). Once again the increased SNR allows the determination of fine details as the shoulders.

A side view of the mouse was also taken (Figure 6.16). Details of the spine and skull (orbits, brain vault, jaw) are clear. The rib cage is also seen under the thoracic spine and two ribs may also have been separated. A hint of the position of the arms also appears. These attributions, however, need further supporting evidence.

In Figure 6.18 it is seen that the resolution of the coded aperture is much superior to that of a

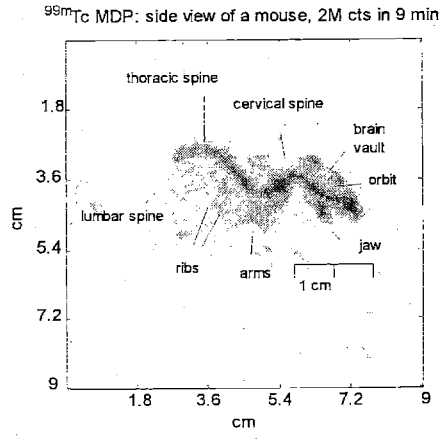


Figure 6.16: bone scan of a mouse: side view. Activity injected: 300 μCi. 521+531 s exposure for $2 \times 2 \times 10^6$ counts.

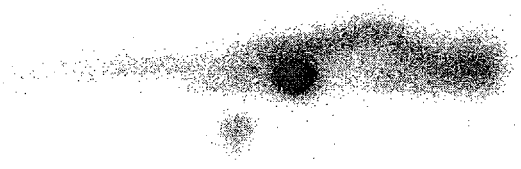


Figure 6.17: comparison to a high-sensitivity collimator. System resolution: 15.2 mm at 10 cm. Courtesy of Siemens Nuclear Medicine Group.

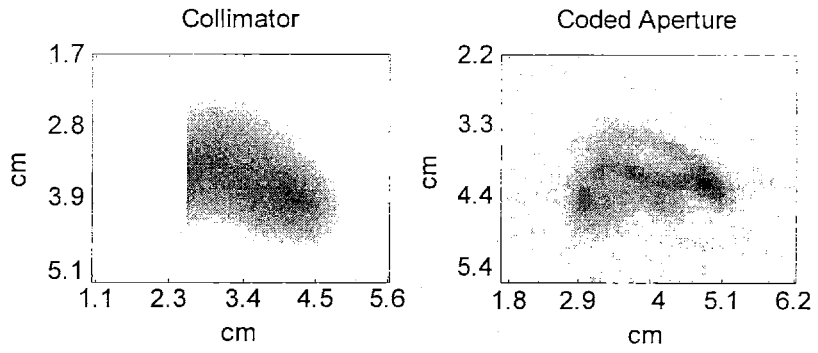


Figure 6.18: comparison of coded aperture vs. high-sensitivity collimator, for the same activity (1 mCi injected) and exposure time. Activity shielded below neck (data cut in the case of the collimator). Poor performance as compared to previous experiments indicated a suspect bad injection. Code aperture image: 4×10^6 counts total in 26min 25s, corresponding to 1320 cps. Collimator image: 264 cps, after decay correction.

collimator. However, the exposure time is longer than that of the mouse of Figure 6.16, which was injected with about one third of the activity.

The suspicion of a wrong injection was strengthened when a picture of the upper torso was taken (Figure 6.19). This shows a line of activity at the edge of the lead shield. A whole body scan (Figure 6.20) revealed that the bulk of the activity was in the perineum, around the injection point. Not only this explained the poor pick up, but also the activity at the edge of the lead shielding. In fact, the lead screen used to shield activity in the bladder kept the waist of the mouse above its shoulders, so activity loose in the perineum may have can flowed by gravity along the body, up to the diaphragm.

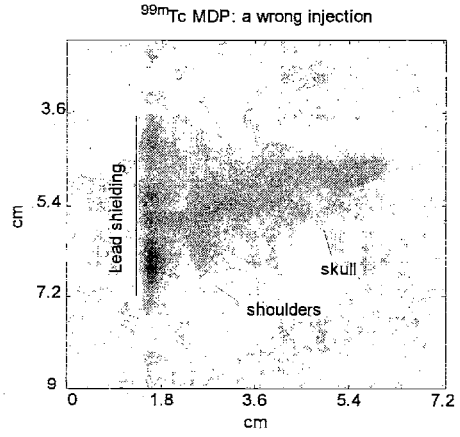


Figure 6.19: scan of a mouse above the chest. Case of wrong injection.

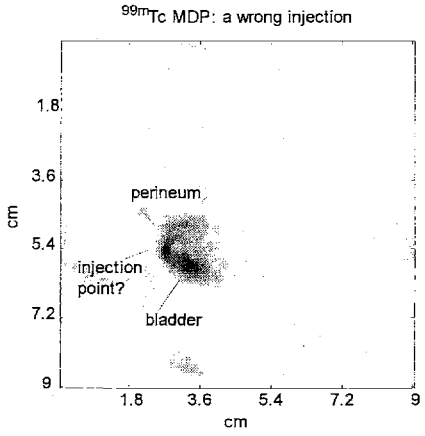


Figure 6.20: whole body of the mouse of Figure 6.19 showing activity loose in the perineum.

6.5 Summary

From the characteristcs of an existing Anger camera and the SNR and artifact reduction ideas developed in previous Chapters a prototype mask was designed and built. Its characteristics are in Table 6.1 and are compared to those of pinholes and collimators in Table 6.2. The geometric resolution of the coded aperture is 1.45 mm. In Chapter 7 is developed the theory necessary to evaluate system resolution, which is shown to be 1.66 mm, which is consistent with experimental evidence. These numbers are lower than the intrinsic PSF of the Anger camera because the camera operates with a relatively high magnification ($m = 4.3$).

The sensitivity of the instrument is about 10 times higher than that of the most sensitive collimator and 100 times higher than that of a pinhole larger than the holes of the mask. However this does not translate immediately in an SNR advantage because of multiplexing. Different resolution should also be considered in a fair comparison.

A picture of the protoype mask is in Figure 6.21. Experiments showed that this mask is capable of high-resolution 2d imaging with very little artifacts. Not all images, however, have the same quality. In agreement with theory, best performance was obtained for sources concentrated on a limited part of the field of view.

Optics (^{99m}Tc)	Collimator		Pinhole	Coded Aperture
	High Sensitivity	Ultra-high-resolution		
Hole diameter (mm)	2.54	1.16	4	1.11
Geometric resolution @ 10 cm (mm)	14.6	4.6	6.2	1.45
System resolution @ 10 cm (mm)	15.2	6.3	6.6	1.67
Sensitivity @ 10 cm (cpm/μCi)	1063	100	123	~10000

Table 6.2: comparison of the characteristics of the coded aperture with conventional optics. Pinhole cone length: 20 cm. Collimators have hexagonal holes.

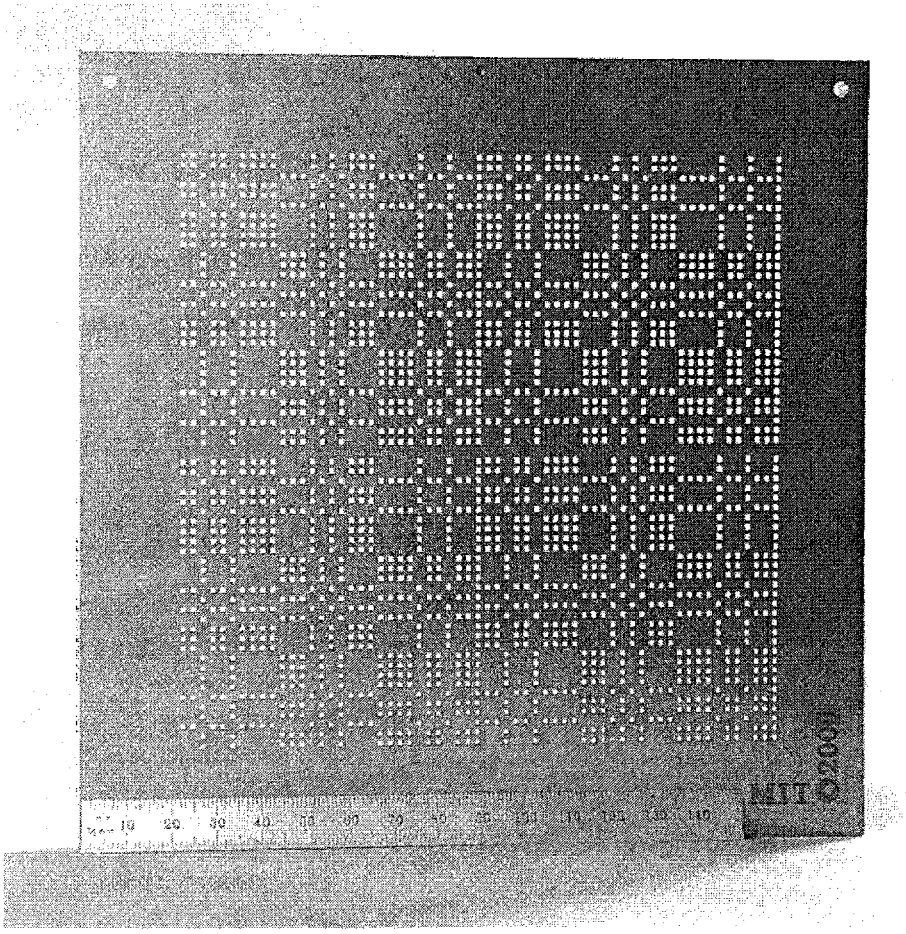


Figure 6.21: the prototype mask. It is a 62×62 mosaicked, pattern centered, anti-symmetric NTHT MURA with open fraction 12.5%. It is made of 12 layers of photo-etched tungsten for a 1.5-mm thickness. Square straight holes of side 1.114 mm.

PART III:

ADVANCED TOPICS AND FUTURE DEVELOPMENTS

Chapter 7 ADVANCED MASK DESIGN

This Chapter gathers two advanced topics: ultimate resolution limit and a more thorough investigation of optimal mask thickness.

In the first section the effect of detector non-idealities is investigated to calculate the system resolution of a coded aperture camera and understand what sets the ultimate limit on system resolution. In designing the prototype mask the acquisition parameters of the E-Cam were assumed fixed. Here different camera configurations are analyzed in the design of a mask designed for the highest possible resolution. Since maximum mask thickness turns out to be a major factor setting the resolution limit, a theory more accurate than the one developed in Chapter 6 was needed. It led to the conclusion that the prototype mask is indeed too thick for ^{99m}Tc but would be about optimal for slightly higher-energy isotopes as ^{111}In. Experimental PSF substantiating this claim close the Chapter.

7.1 Resolution in a real detector

In section 2.6 the resolution of a coded aperture camera was shown to depend on magnification, detector dimensions and the field of view, but the ultimate limit seemed to be set by the mask pixel size p_m. With those arguments, there is no lower limit to resolution, provided mask holes can be made small enough. This would not even reduce the signal throughput, because the total open area can be kept constant by making the coded aperture pattern larger. This surprising result comes from a mix of assumptions on detector, mask and exposure time (or activity): the calculation of resolution has so far assumed an ideal detector, i.e. it has been a calculation of geometric, not system resolution.

In reality, a mask with very small holes would have two problems. The first is that the intrinsic PSF of the detector or the number of pixels in the acquisition may not allow one to resolve the shape of the shadow. The second is that small holes also imply a reduced thickness to avoid collimation and ensuing artifacts. In the following a general theory to calculate system resolution is developed and then applied to the case of the prototype mask.

7.1.1 Mathematical derivation of system resolution

The main non-idealities of an Anger camera are two: intrinsic PSF and discrete data handling.

The intrinsic PSF of the Siemens E-Cam is Gaussian, with a FWHM of 3.7 mm at its center, increasing to 3.9 mm at its sides (see §3.2.2 and eq. (3.12)). The shape of the projection of the mask as seen by the phototubes is the ideal, continuous projection of eq. (2.18) convolved with PSF_i:

$$R(\vec{r}) \propto (O' * A') * PSF_i \tag{7.1}$$

where A' is the geometric projection of the mask cast on the detector by a point source. To keep the formulation general, A' is defined in terms of the finely sampled array $\mathbf{A}^F$ (see §2.7.4) and the mask hole shape S'_m, which gives the shape of the projection of one (finely sampled) mask position on the detector:

$$A'(\vec{r}^{\,i}) = A'_\delta * S'_m = \sum_{i,j} \mathbf{A}^F_{i,j} \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) * S'_m = \sum_{i,j} \mathbf{A}^{F\delta}_{i,j} * \mathbf{H}_\alpha \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) * S'_m \tag{7.2}$$

$\mathbf{H}_\alpha$ is an $\alpha \times \alpha$ square array of 1s. The continuos distribution R is then digitized in pixel counts. This process is equivalent to integrating over the area of a pixel and then sampling on a grid with pitch equal to the pixel size p_d. This was already encountered in eq. (2.67), which is repeated here for convenience:

$$\mathbf{R}_{r,s} = \iint_{\vec{r}} R(\vec{r}) S_p(\vec{r} - \vec{r}_{r,s}) d^2\vec{r} \tag{2.67}$$

or, shortly,

$$\mathbf{R} = R \times S_p|_{r,s} \tag{7.3}$$

where the vertical bar indicates evaluation at point $\vec{r}_{r,s}$. This is the sampling grid and may or may not be centered on or have the same pitch as the grid $\vec{r}^{\,i}_{i,j}$ of the centers of the projections of the pinholes of the mask. If the pitch is the same, then a mask pinhole is projected on an integer number of detector pixels and α is integer. In this case, there are no defocusing artifacts (see §5.2.2). If the two grids are also superimposed there are no border artifacts (see §5.2.1).

The continuous function R has turned in to the discrete array **R**. In practice, this is the data generated by the Anger camera and fed to the decoding algorithm to give the reconstructed image:

$$\hat{\mathbf{O}} = \mathbf{R} \otimes \mathbf{G} \tag{7.4}$$

The analytical calculation is more convenient if the continuous function $R \times S_p$ is correlated with G_δ, the continuous analogue of **G** (see eq. (2.82)) and the result sampled:

$$\hat{\mathbf{O}} = R \times S_p \otimes G_\delta \,|_{r,s} \tag{7.5}$$

This derivation is different from that of section 2.6 because of the additional convolution with PSF_i and S_p. Substituting in eq. (7.1), with the help of eq. (7.2):

$$\hat{\mathbf{O}} = O' * A'_\delta * S'_m * PSF_i \times S_p \otimes G_d \Big|_{r,s} = O' * A'_\delta \otimes G_d * S'_m \times S_p * PSF_i \,|_{r,s} \tag{7.6}$$

Like A'_δ, G_δ can be written in terms of the finely sampled array $\mathbf{G}^F$:

$$G_\delta = \sum_{i,j} \mathbf{G}^F_{i,j} \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) = \sum_{i,j} \mathbf{G}^{F\delta}_{i,j} * \mathbf{H}_\beta \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) \tag{7.7}$$

where $\mathbf{H}_\beta$ is a $\beta \times \beta$ square array of 1s and $\mathbf{G}^{F\delta}$ was defined in §2.7.4. β can range from 1, in which case $\mathbf{G} = \mathbf{G}^{F\delta}$ (δ decoding), to α. Substituting eq. (7.2) and eq. (7.7) in eq. (7.6) gives:

$$\begin{aligned}
\hat{\mathbf{O}} &= O' * \sum_{i,j} \sum_{i+k,j+l} \sum_{i,j} \mathbf{A}^{F\delta}_{i,j} \mathbf{G}^{F\delta}_{i+k,j+l} * \mathbf{H}_\alpha * \mathbf{H}_\beta \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i+k,j+l}) * S'_m \times S_p * PSF_i \Big|_{r,s} = \\
&= O' * \sum_{i,j} \sum_{i,j} \mathbf{H}_\alpha * \mathbf{H}_\beta \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) * S'_m \times S_p * PSF \Big|_{r,s} = \\
&= O' * \left(\sum_{i,j} \mathbf{H}_\alpha \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) * S'_m \right) * \left(\sum_{i,j} \mathbf{H}_\beta \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) \times S_p \right) * PSF \Big|_{r,s}
\end{aligned} \tag{7.8}$$

where:

$$\sum_{i,j} \mathbf{A}^{F\delta}_{i,j} \mathbf{G}^{F\delta}_{i+k,j+l} = \delta(k,l) \tag{7.9}$$

was used with the fact that the grids associated to **A** and **G** have the same pitch. In the correlation:

$$\sum_{i,j} \mathbf{H}_\alpha \delta(\vec{r}^{\,i} - \vec{r}^{\,i}_{i,j}) * S'_m \tag{7.10}$$

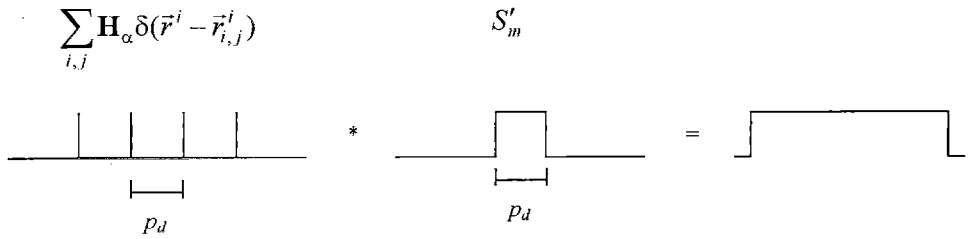


Figure 7.1: convolution of a grid of δ functions with a rect function having the same side as the grid spacing.

the first factor is an infinite grid of δ's, only an $\alpha \times \alpha$ square of which has a non-zero coefficient. In the case of totally open square mask holes, S'_m is a square having exactly the size of the distance between the δs of the first factor. As shown in Figure 7.1, the result is that the convolution is an $\alpha p_d \times \alpha p_d$ rectangle. Similarly, S_p generates a $\beta p_d \times \beta p_d$ square. With the symbol:

$$\mathrm{rect}(\vec{v};c) = \begin{cases} 1 & -c/2 \le \vec{v}_x \le c/2 \ \text{ and } \ -c/2 \le \vec{v}_y \le c/2 \\ 0 & \text{elsewhere} \end{cases} \tag{7.11}$$

$\hat{\mathbf{O}}$ becomes:

$$\hat{\mathbf{O}} = O' * \mathrm{rect}(\vec{r}^{\,i};\alpha d_p) * \mathrm{rect}(\vec{r}^{\,i};\beta d_p) * PSF_i|_{r,s} \tag{7.12}$$

The system PSF, which is the basis for the evaluation of the resolution, is calculated substituting O' with a point source:

$$PSF_{sys} = \delta * \mathrm{rect}(\vec{r}^{\,i};\alpha d_p) * \mathrm{rect}(\vec{r}^{\,i};\beta d_p) * PSF_i = \mathrm{rect}(\vec{r}^{\,i};\alpha d_p) * \mathrm{rect}(\vec{r}^{\,i};\beta d_p) * PSF_i|_{r,s} \tag{7.13}$$

The system PSF is the convolution of two rectangles, one of the size of the projection of the mask pixels and the other of the size of the detector pixels, with the intrinsic PSF.

7.1.2 Analysis of the coded aperture PSF

For simplicity, the issue of sampling is momentarily set aside, and the result analyzed in 1d. With this simplification the system resolution λ_s is the FWHM of PSF_i rescaled in terms of object space (see section 2.6):

$$\lambda_s = \frac{FWHM_{PSF_{sys}}}{m-1} \tag{7.14}$$

The convolution of two rectangles is a trapezoid. Since one of the rectangles has side $\alpha\, p_d$, the other $\beta\, p_d$ and $\beta \le \alpha$, the trapezoid has area $\alpha\beta\, p_d$ and its FWHM is $\alpha\, p_d$ (Figure 7.2). The remaining convolution with PSF_i can be split in the sum of three parts, each being the convolution of a straight line with a Gaussian. For a the generic line:

$$y = mx + q \tag{7.15}$$

the convolution between x_i and x_f is:

$$\int_{x-x_i}^{x-x_f} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} (m(x-\xi)+q)d\xi \tag{7.16}$$

which, with a little math, becomes:

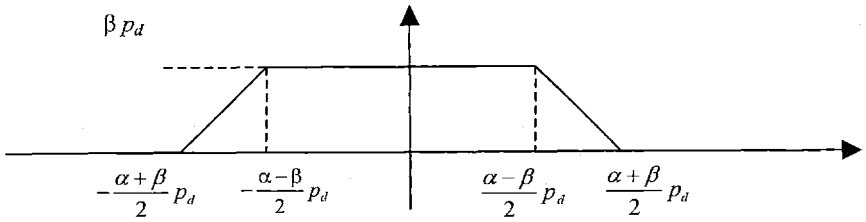


Figure 7.2: trapezoid coming from the convolution of two rectangles of side αp_d and βp_d.

Interval $[x_i, x_f]$	Interval (with dimensionless par.)	Equation	m	q
$[-\frac{\alpha+\beta}{2}p_d, -\frac{\alpha-\beta}{2}p_d]$	$[-\frac{1+r}{2}, -\frac{1-r}{2}]\, s\sqrt{2}\sigma$	$y = x + \frac{\alpha+\beta}{2}p_d$	1	$\frac{\alpha+\beta}{2}p_d$
$[-\frac{\alpha-\beta}{2}p_d, \frac{\alpha-\beta}{2}p_d]$	$[-\frac{1-r}{2}, \frac{1-r}{2}]\, s\sqrt{2}\sigma$	$y = \beta\, p_d$	0	$\beta\, p_d$
$[\frac{\alpha-\beta}{2}p_d, \frac{\alpha+\beta}{2}p_d]$	$[\frac{1-r}{2}, \frac{1+r}{2}]\, s\sqrt{2}\sigma$	$y = -x + \frac{\alpha+\beta}{2}p_d$	-1	$\frac{\alpha+\beta}{2}p_d$
$s = \frac{\alpha p_d}{\sqrt{2}\sigma}, \; r = \frac{\beta}{\alpha}$				

Table 7.1: parameters useful in the calculation of the system PSF.

$$\sigma\left\{\frac{mx+q}{2\sigma}\left[erf\left(\frac{x-x_f}{\sqrt{2}\sigma}\right)-erf\left(\frac{x-x_i}{\sqrt{2}\sigma}\right)\right]+\frac{m}{\sqrt{2\pi}}\left(e^{\frac{(x-x_f)^2}{2\sigma^2}}-e^{\frac{(x-x_i)^2}{2\sigma^2}}\right)\right\} \tag{7.17}$$

where the error function is:

$$erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{7.18}$$

All that is left is to find the parameters m and q for each of the three parts of the trapezoid. From simple geometry the results of Figure 7.2 and Table 7.1 were obtained. Substitution of these values in eq. (7.17) allows the calculation of the three terms whose sum is the PSF of the system. The algebra is straightforward and not followed by major simplifications, so that it makes little sense to report the result here. Sample plots of the system PSF are provided in Figure 7.3 instead. For increasing α, the PSF gradually deviates from Gaussian and takes the form of a trapezoid with smoothed edges. In this process the FWHM increases, because the resolution of the coded aperture (the geometric resolution) is now added to the intrinsic resolution of the detector. Conversely, decreasing α improves geometric resolution until intrinsic resolution dominates.

PSF_{sys} is a function of the variable x only and depends directly on σ and through the boundaries x_i and x_f of Table 7.1 on the three parameters α, β and p_d. One parameter can be eliminated because x_i and x_f are actually a function of two parameters only, for example the product αp_d and $r=\frac{\beta}{\alpha}$. With the

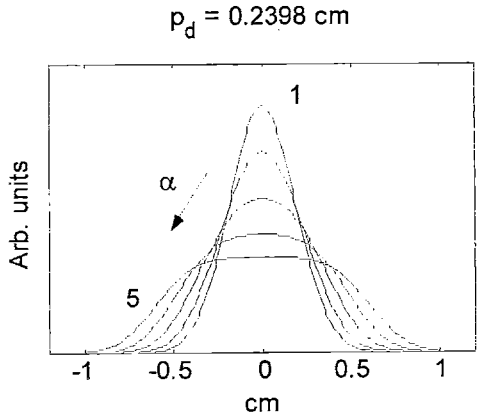


Figure 7.3: system PSF for different values of α. σ = 1.571 mm, β = 1. As α increases the intrinsic PSF is broadened by the projection of the mask holes.

definition of the dimensionless variables $\xi = \frac{x}{\sqrt{2}\sigma}$ and $s = \frac{\alpha p_d}{\sqrt{2}\sigma}$, the dependence of PSF_{sys} on σ is also removed, because σ becomes simply a scaling constant.

In conclusion, PSF_{sys} depends only on a variable, ξ, and two parameters, s and r, which are the only ones to influence its FWHM. s is proportional to the ratio of the projection of a mask hole (through eq. (2.96)), which is strictly related to the geometric resolution (see section 2.6), to a measure of the intrinsic resolution of the detector. The parameter r is uniquely related to the decoding strategy. The parameters r and s can vary for a number of reasons: because p_d or σ change, or because α changes because magnification changes (affecting r as well).

To quantify the deviation of the system resolution from the geometric resolution given by the eq. (2.87), was defined the factor χ:

$$\lambda_s = \lambda_g \chi = p_m \frac{m}{m-1} \chi \tag{7.19}$$

where χ takes the dependence on s and r:

$$\chi = \chi(s, r) \tag{7.20}$$

χ must be calculated numerically, but its dependence on the parameters can be predicted with some intuitive arguments. From the physical meaning of s, if geometric resolution is poor as compared to intrinsic resolution, s is high. The effects of intrinsic resolution are too small to be seen and resolution

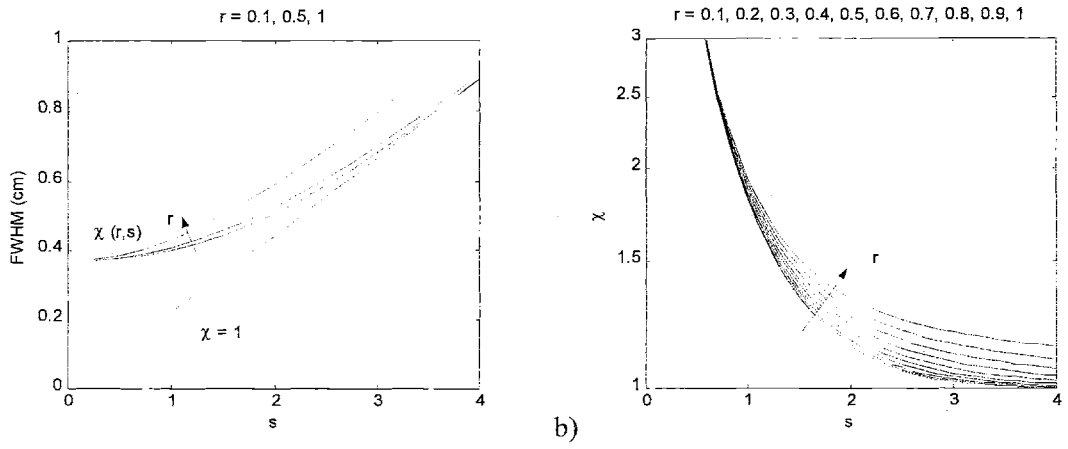


Figure 7.4: a) FWHM as a function of s for different values of r. The line $\chi = 1$ indicates the hypothetical case of a detector with $\sigma = 0$. b) χ as a function of s and r.

must essentially equal geometric resolution: in the limit χ approaches 1. As s decreases, the effect of intrinsic resolution becomes visible. In Figure 7.4a one can see that despite the decrease in s, the FWHM does not get below 3.7 mm, the intrinsic FWHM. Correspondingly, χ rises from 1, indicating that the hypothesis of ideal detector does not hold. Deviation from geometric resolution makes χ larger than 1. The first lesson learned is the s should be kept as small as possible, but improvements have a limit and decreases in s eventually become too costly for little additional advantage.

In Figure 7.4b is shown the dependence of χ on r and s. Low values of r (δ decoding) give better resolution but caution should be used in determining the minimum r. In fact, if a given configuration enjoys low s, it may be due to a low α. This increases the minimum r that can be used, because r is the ratio β / α, with β integer greater or equal to 1. A common case is $\alpha = 2$, in which case r is no less than 0.5. This is also an example that not all configurations with the same product $\alpha\, p_d$ are equivalent, as one may think because of the same s: for the same product, the configuration with maximum α (and minimum p_d) should be preferred because r can be made smaller. In the next paragraph is discussed a second reason leading to the same conclusion.

7.1.3 Effects of discretization on resolution

The issue of the system PSF actually being a sampled version of a continuous function was so far neglected. Discretization may degrade resolution a little more because point sources that would be resolved if the image were continuous may not be resolved after sampling of the image. An example is that of two Gaussian curves separated by a FWHM: such curves show a 7% dip between the peaks (see Figure 7.5). If the image is sampled too coarsely, samples may be taken just at the peaks, and separation

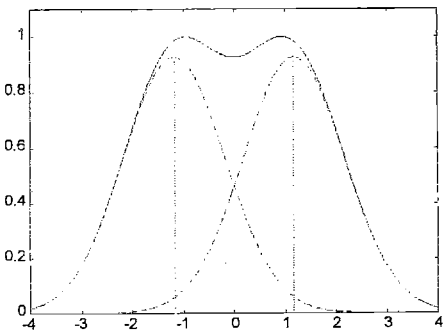


Figure 7.5: two Gaussian curves with $\sigma = 1$, displaced by a FWHM (2.3548) and summed give a curve with a minimum at 0.9272.

would be lost.

A first solution is to use a small sampling period p_d, so that dips will not be missed, but finer sampling may hurt the SNR of the image. This loss, however, can be compensated by increasing β to keep summing over the same detector area in reconstruction. To show this with an example, if p_d is halved, α is automatically doubled if mask and geometry are constant because a mask pinhole is projected on twice as many pixels (in 1d). s remains constant and, if δ decoding is adopted, reconstruction includes data from a smaller pixel only, i.e. must rely on a smaller number of counts, which reduces the SNR. But if β is also doubled, i.e. set to 2, an area of 2×2 pixels, equivalent to the old p_d, contributes data, leaving the SNR intact. r also remains constant, so the resolution indicated by eq. (7.19) is constant, but resolution of the real image is a little better because of the finer pixels. An image completely equivalent to the previous can be obtained by discarding every other column and row of this new image.

The lesson learned by looking at sampling is that for a given s, which may be fixed because of the geometry and the mask, a combination that minimizes p_d is to be preferred. Better resolution is obtained because of better sampling. In the next section is shown that pixels have a typical size of 0.6 mm or less. This number is 6 times less than the intrinsic FWHM. In these conditions, we can be confident that sampling does not cause additional resolution losses and shall be happy with the resolution theory developed for the continuous version of the system PSF. To maintain the SNR, the data so collected can always be clustered in larger pixels, giving images equivalent to those obtained with larger pixels, a process that can be incorporated in decoding using the parameter β.

7.1.4 System resolution of the prototype coded aperture camera

For the coded aperture designed in Chapter 6, $\alpha = 2$, $p_d = 2.398$ mm and $\beta = 1$. These data imply $s = 2.1546$ and $r = 0.5$. With these values $\chi = 1.146$. From eq. (7.19) the system resolution of the coded aperture camera used is:

$$\lambda_s = p_m \frac{m}{m-1} \chi = 1.66 \text{ mm} \tag{7.21}$$

If conventional instead of δ decoding is used, then $r = 1$, $\chi = 1.297$ and $\lambda_g = 1.88$ mm. This value should be compared to the Siemens high-sensitivity low-energy (^{99m}Tc) collimator, whose system resolution is 15.2 mm, the ultra-high resolution collimator (6.3 mm), and the 4 mm pinhole (6.6 mm) (see Table 6.2).

7.2 Choosing a configuration

In Chapter 6 the configuration of the E-Cam (the way it makes pixels) was set for simplicity to a commonly used one. Better performance can be obtained by optimizing its setup.

The detector of the Siemens E-Cam has an active area of 53.3 × 38.7 cm. Pixels are made by dividing a nominal area of 61.4 × 61.4 cm in 32 × 32, 64 × 64, 124 × 124, 256 × 256, 512 × 512 or 1024 × 1024 pixels. Only as many pixels as fit in the active area are actually active. For example, for 256 × 256 pixels, each pixel is 61.4 cm / 256 = 2.398 mm and only 38.7 / 0.2398 = 161 pixels are active (222 along the other direction). Smaller pixels can be obtained by using a zoom factor, which reduces the nominal area by a factor of 1, 1.23, 1.45, 1.78, 2, 2.29, 2.67 or 3.2. For a zoom of 1.78 the nominal area is 61.4 / 1.78 = 34.5 cm (on both sides), which is less than the active area, so all pixels will be active. In these cases parts of the active area are not used. Since we know that best resolution for a given field of view is obtained with the largest detector available (see section 2.6), it is not beneficial to use this or any higher zoom.

From §7.1.2, the most advantageous configurations are those with the smallest pixel size, so zoom 1.45 and 1024 pixels should be preferred. In this case $p_d = 0.4134$ mm. Also note that the largest pixel size for 1024 pixels is 0.6 mm for zoom 1.

What is the best configuration that can be used with the designed mask? Since the camera must make at least 124 pixels to sample properly its projection, only configurations with 256 pixels or more are acceptable[15]. For each of these, α can be set to its maximum possible integer value, obtained by dividing (along the short side of the detector) the number of detector pixels by the number of mask pixels, i.e. 62. Once a choice of β is made, all parameters needed to calculate the system resolution of the prototype mask with all configurations are available. Depending on how integer numbers fit in the detector's active area, results can be a bit unpredictable and must be calculated one by one. They are reported in Table 7.2. Two decoding strategies are considered. To compare configurations at constant SNR $\beta = \alpha$ should be used while $\beta = 1$ gives best resolution regardless of SNR. The result is the same in both scenarios: zoom 1.45 and minimum p_d should be used. Differences are not dramatic. By using this configuration in place of that of Chapter 6 system resolution improves from 1.66 mm to 1.4 mm for δ decoding and from 1.88 mm to 1.65 mm for conventional decoding. Nonetheless, the advantage comes only at the expense of having to handle larger files.

---

[15] Due to the restriction on the active area, the configuration at 128 pixels actually makes 81, 99 and 117 pixels along the shorter side for zoom 1, 1.23 and 1.45, respectively.

pixels	zoom	p_d (mm)	α	β	χ	λ_s (mm)	β	χ	λ_s (mm)
1024	1.45	0.414	15	1	1.0312	1.400	15	1.2124	1.646
512	1	1.199	5	1	1.0419	1.426	5	1.2215	1.671
1024	1	0.600	10	1	1.0382	1.420	10	1.2215	1.671
256	1.23	1.950	3	1	1.0570	1.454	3	1.2284	1.690
512	1.23	0.975	6	1	1.0457	1.439	6	1.2284	1.690
1024	1.23	0.487	12	1	1.0430	1.435	12	1.2284	1.690
512	1.45	0.827	7	1	1.0471	1.444	7	1.2314	1.699
256	1.78	1.347	4	1	1.0720	1.505	4	1.2537	1.761
512	1.78	0.674	8	1	1.0652	1.496	8	1.2537	1.761
1024	1.78	0.337	16	1	1.0635	1.493	16	1.2537	1.761
256	1.45	1.654	3	1	1.1081	1.592	3	1.2833	1.843
256	1	2.398	2	1	1.1457	1.662	2	1.2968	1.882

Table 7.2: E-Cam configurations for use with the designed coded aperture. Two decoding strategies are considered. To compare configurations at constant SNR use $\beta = \alpha$, while to achieve best resolution $\beta = 1$. The result is the same in both scenarios. Configuration with zoom 1.78 were also included to make sure the effects of using integer numbers did not change the result. $m = \alpha\, p_d / p_m$ is different for every configuration.

7.3 The design of an optimal resolution mask

A very different question is that of incorporating different detector configurations in mask design to optimize resolution. An additional advantage that could be included is that deriving from use of non-integer α (see section 6.3). This avoids the need to have α integer, which forced the use of an area of the detector smaller than its active area, unnecessarily limiting resolution (see 2.6). In this hypothesis m is determined by the design field of view (eq. (2.79)):

$$m = \frac{d_d}{FoV} + 1 = \frac{38.7}{9} + 1 = 5.3 \tag{7.22}$$

From this m and eq. (2.87) one can calculate that p_m is multiplied by a factor of 1.23 to give resolution: reducing the field of view to improve magnification can not improve resolution by more than 23%.

Best resolution is obtained for the minimum s and δ decoding. s is proportional to p_d, whose minimum value (excluding configurations with a smaller detector active area) is 0.4134 mm, obtained for zoom 1.45 and 1024 pixels. For a given configuration, σ and p_d are constants, so minimum s is the same as minimum α,[16] which, from the discussion of border artifacts (see §5.2.1), is 2. With these numbers $s = 0.3722$. For $r = 0.5$, FWHM = 3.754 mm and $\chi = 4.5388$, while for $r = 1$, FWHM = 3.785 mm and $\chi = 4.5771$. The value of the FWHM indicates that the limit on resolution is being set by the intrinsic PSF. The value of χ indicates that system resolution is 4.5 times worse than geometric resolution, a fact suggesting further investigation.

The calculation can be carried on to calculate all mask features. Since the number of active pixels is $38.7 / 0.04134 = 936$, for $\alpha = 2$ the mask array is 468×468[17]. System resolution is then given by eq. (7.14): $\lambda_s = 0.3754 / 4.3 = 0.087$ cm or 0.87 mm. From eq. (2.96) the mask pixel size is $p_m = 2 \times 0.0413 / 5.3 = 0.015$ cm or 150 μm. While technology does exist to fabricate such small holes, the question is whether it is worth it, given the great loss caused by the intrinsic PSF. For example, a similar calculation shows that with zoom 1, 1024 pixels ($p_d = 0.599$ mm), $\alpha = 4$ and $\beta = 4$, then $\chi = 1.8365$, FWHM = 4.405 mm and $\lambda_s = 1.024$ mm. Resolution is 17% worse but mask holes are $p_m = 4 \times 0.599 / 5.3 = 0.452$ mm, which is almost three times as large. Holes of this size would also allow thicker masks, and, thus, higher SNR.

The best approach is to scan systematically all detector configurations and then compare mask designs. Of course, the procedure can be automated. For a given detector configuration a minimum acceptable value of α (at least 2) is set and the corresponding maximum number of pixels of the mask calculated, with the restriction that the pattern must be an anti-symmetric NTHT MURA. When this number is known, α can be recalculated as the ratio of detector pixels to mask pixels. As $m = 5.3$, eq. (2.96) gives p_m:

$$p_m = \frac{\alpha p_d}{m} \tag{7.23}$$

which can be substituted with eq. (7.22) in the expression for system resolution eq. (7.19) to get:

---

[16] Setting α to its minimum value is not in contradiction with what was done in the previous section, where the mask already existed and its performance had to be optimized. The first concern is to minimize s, i.e. p_d and α. In the previous section s was constrained because the mask was already built. In these conditions, where the product of α and p_d is constrained, α should be maximized.

[17] Such array may not exist and if it does is not an NTHT MURA because $468 / 2 = 234$ which is not prime. This does not invalidate the discussion following.

$$\lambda_s = \frac{\alpha d_p}{m}\frac{m}{m-1}\chi(r,s) = \alpha d_p \frac{FoV}{d_d}\chi(r,s) = \sqrt{2}\,\sigma\, s \frac{FoV}{d_d}\chi(r,s) \tag{7.24}$$

which is a function of r and s only, i.e. of the detector configuration and the choice of α and β. Since the scope is to achieve maximum resolution, $\beta = 1$. All is known to calculate λ_s for the configuration at hand. The procedure can then be repeated for all detector configurations and minimum values of α.

α_{min}	zoom	α	n	p_m (mm)	χ	λ_s (mm)
2	1.4500	2.0595	454	0.161	4.4108	0.874
2	1.2300	2.0759	382	0.191	3.7344	0.880
3	1.4500	3.0960	302	0.242	2.9791	0.888
2	1	2.1358	302	0.242	2.9874	0.890
3	1.2300	3.0267	262	0.279	2.6111	0.897
4	1.4500	4.3692	214	0.341	2.1690	0.912
3	1	3.0140	214	0.341	2.1747	0.915
5	1.4500	5.6325	166	0.440	1.7435	0.945
4	1.2300	4.7771	166	0.440	1.7454	0.946
5	1.2300	5.0190	158	0.462	1.6766	0.955
4	1	4.0823	158	0.462	1.6788	0.956
6	1.4500	6.5845	142	0.514	1.5404	0.976
6	1.2300	6.7203	118	0.619	1.3510	1.030
7	1.4500	7.9237	118	0.619	1.3499	1.030
5	1	5.4661	118	0.619	1.3525	1.032
8	1.4500	9.9468	94	0.777	1.1817	1.131
7	1.2300	8.4362	94	0.777	1.1825	1.132
8	1.2300	8.4362	94	0.777	1.1825	1.132
6	1	6.8617	94	0.777	1.1835	1.133
7	1	7.5000	86	0.849	1.1353	1.188
8	1	10.4032	62	1.178	1.0308	1.496

Table 7.3: system resolution for maximum usage of the detector. All results for zoom 1.78 resulted in worse resolution with smaller mask pixels and were not included. m is the same for all configurations.

Results are in

Table 7.3. The best configuration achieves $\lambda_s = 0.874$ mm with a 454 × 454 pattern. At the other end is the familiar 62 × 62 mask, now achieving a better resolution than previously encountered because of the relaxation of the condition on α. It is difficult to set a threshold on the minimum acceptable mask pinhole size. An attractive pattern was the 158 × 158 because it provides sub-millimeter system resolution and needs values of α not too distant from an integer value, which could result in better results of the decoding algorithm. The problem is again to optimize the thickness and then verify if, for the resulting transparency, the SNR is acceptable. Of course, the SNR depends on activity, but this effect is only partially included in the treatment done in terms of the figure of merit. A more quantitative approach is needed.

7.4 An improved figure of merit

Low transparency has two beneficial effects on the SNR: the net signal is higher and the sidelobe variance lower. From Table 4.1, the SNR of a NTHT (M)URA mask with an open fraction of 1 / 8, in the assumption of no background, depends on t as follows (see also Figure 7.6):

$$\frac{SNR(t)}{SNR(t=0)} = \frac{1-t}{\sqrt{1+7t}} \tag{7.25}$$

From this perspective, optimization of the SNR requires a mask as thick as possible, but this

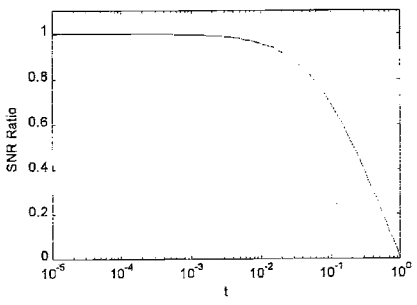


Figure 7.6: SNR loss from an ideally opaque mask as a function of transparency t.

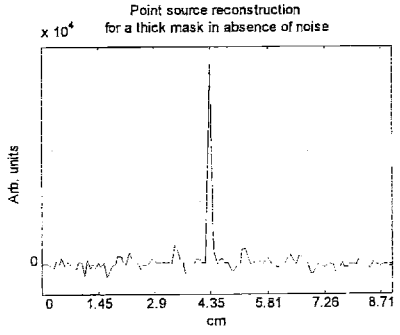


Figure 7.7: reconstruction of a point source with no statistical noise. Variation on the side values is due to thickness artifacts, not to randomness. This is revealed by the approximate symmetry of the peaks.

generates thickness artifacts. As already pointed out, on a slice through the reconstructed peak thickness artifacts seem to have no particular structure: deviations from ideality resemble deviations due to low counting statistics. Of course, there is a big difference between the two effects: thickness artifacts are deterministic and do not depend on exposure time or object activity. However, it is not too far from truth to say that artifacts influence the image very much like poor statistics and, in this sense, can be quantified in terms of the SNR. In this way the problem of optimizing mask thickness is reduced to that of finding an optimal value of some combined SNR as a function of mask thickness. An overall figure of merit FoM_t was defined. It is made of two parts.

7.4.1 Artifact contribution

The first contribution to FoM_t is its artifact part FoM_a, which can be quantified by simulating the reconstruction of a point source in absence of noise. A typical result is in Figure 7.7. FoM_a is given by the ratio of a signal, the expected height of the peak $NI_T(1-t)$, to the average standard deviation from 0 of the sidelobes:

$$FoM_a = \frac{NI_T(1-t)}{\sigma_a} \tag{7.26}$$

where σ_a is obtained with a simulation in absence of random noise.

7.4.2 Statistical contribution

The second piece of FoM_t, the SNR deriving from Poisson noise, can, in principle, be taken directly from the SNR theory of Chapter 4. Unfortunately, predictions would not be accurate in this case because that derivation does not account for the effect of the intrinsic PSF. A precise calculation must restart from the expression of the noise (Poisson, hence the subscript p) variance valid for an NTHT (M)URA:

$$\sigma^2_{p\,i,j} = I_T\left[(1-t)\sum_{k,l}\psi_{k,l}\sum_{u,v}\mathbf{A}_{k,l}\mathbf{G}^2_{i,j} + tN_T^0\right] \tag{7.27}$$

but, for a NTHT (M)URA mask with $e = 2$, on the j^{th} line:

$$\sum_{u,v} \mathbf{A}_{k,l}\mathbf{G}^2_{i,j} = \begin{cases} 0 & \text{in 1/2 cases} \\ N & \text{in 1/2 cases} \end{cases} \tag{7.28}$$

Due to the intrinsic PSF, even for a point-like source ψ_{ij} is spread over a few pixels. In our conditions, the projection of a mask hole is about 4.8 mm (2.4 mm for sub-millimeter resolution designs), while the FWHM of the intrinsic PSF is 3.7 mm. This means that the central resolution element contains only about half of the activity, while the remaining is divided between the two first neighbors. With this:

$$\sum_{k,l}^{N_T} \boldsymbol{\Psi}_{k,l} \sum_{u,v} \mathbf{A}_{k,l}\mathbf{G}^2_{i,j} = \left\{ \begin{array}{ll} \frac{1}{2}N + \left(\frac{1}{4}+\frac{1}{4}\right)0 & \text{for } j \text{ even (odd)} \\ \frac{1}{2}0 + \left(\frac{1}{4}+\frac{1}{4}\right)N & \text{for } j \text{ odd (even)} \end{array} \right\} = \frac{1}{2}N \tag{7.29}$$

Fortunately, the case of reconstruction of the peak need not be separated from the reconstruction of adjacent positions. The fluctuation in both cases is:

$$\sigma^2_{p\,i,j} = I_T\left[(1-t)\frac{1}{2}N + tN_T^0\right] = I_T\frac{N}{2}\left[1-t+4t\right] = I_T\frac{N}{2}\left[1+3t\right] \tag{7.30}$$

and the SNR:

$$SNR = \frac{I_T\frac{N}{2}(1-t)}{\sigma_{p\,i,j}} = \frac{I_T\frac{N}{2}(1-t)}{\sqrt{I_T\frac{N}{2}\left[1+3t\right]}} = \sqrt{I_T\frac{N}{2}}\frac{(1-t)}{\sqrt{1+3t}} \tag{7.31}$$

where the signal at the peak was divided by two because of the intrinsic PSF.

This is the part of the FoM associated to Poisson noise, FoM_p:

$$FoM_p = \frac{NI_T(1-t)}{2\sigma_{p\,i,j}} = \sqrt{I_T\frac{N}{2}}\frac{(1-t)}{\sqrt{1+3t}} \tag{7.32}$$

This expression also shows that if the signal is rescaled to $NI_T(1\text{-}t)$, the appropriate standard deviation is $2\sigma_p$.

7.4.3 *Aggregate figure of merit*

FoM_p and FoM_a must now be combined in FoM_t. Defining the signal as $NI_T(1-t)$ and assuming that the two sources of noise are not correlated, FoM_t is defined as:

$$FoM_t = \frac{NI_T(1-t)}{\sqrt{\sigma_a^2 + 4\sigma_p^2}} \tag{7.33}$$

which leads to:

$$FoM_t = \frac{1}{\sqrt{\frac{1}{FoM_a^2} + \frac{1}{FoM_p^2}}} = \frac{1}{\sqrt{\frac{1}{FoM_a^2} + \frac{2}{NI_T}\frac{1+3t}{(1-t)^2}}} \tag{7.34}$$

where care must be taken that in I_T are counted only photons that actually reach the detector. These are less than predicted by looking at the solid angle subtended by all holes in the ideal mask because of collimation caused by mask thickness. This reduction can be handled by calculating a collimation factor with a noiseless simulation.

7.4.4 *Determination of mask thickness*

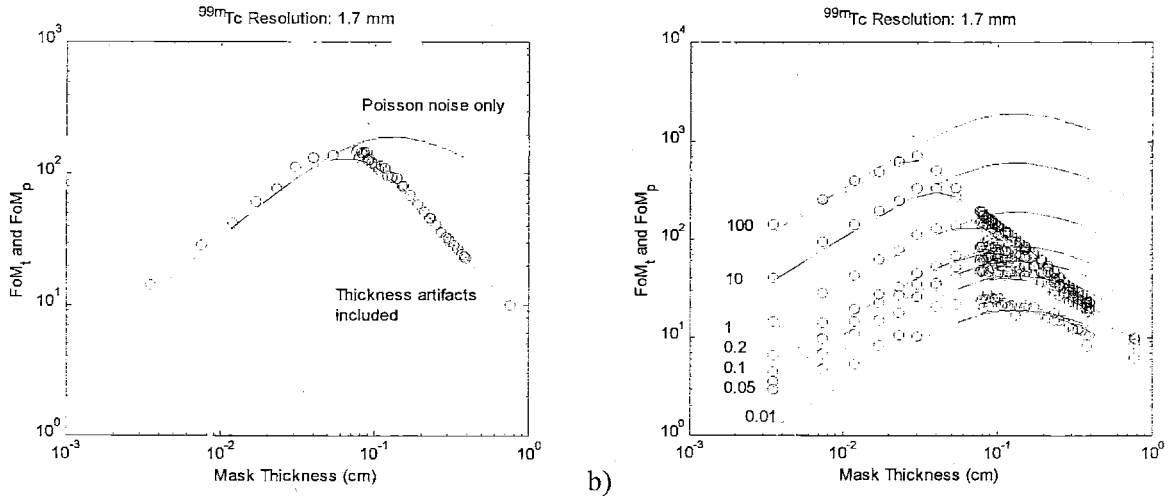


Figure 7.8: a) The top curve is the theoretical prediction of FoM_p. The other curve is the theoretical prediction of FoM_t, which also includes thickness artifacts. Dots indicate the result of simulations with noise and are in agreement with the theoretical prediction. Exposure time 8000 s, activity 0.1 μCi. b) Same as a), for different time. The number next to each set of curves indicates 0.1 μCi of activity for 8000 seconds, i.e. 29.6 million decays.

An example of FoM_t as a function of t is in Figure 7.8a. Two curves are drawn, FoM_t and FoM_p. FoM_p is a theoretical prediction of the SNR due to Poisson noise, taking into account the reduction in the count rate due to collimation (obtained with a simulation with no statistical noise), but does not include the effect of thickness artifacts. This was estimated using the same noiseless simulation to get FoM_t (lower curve). The dots show FoM_t as calculated directly from the reconstruction of simulations with noise. Figure 7.8a shows good agreement. Since FoM_p accounts for the reduction in counts due to mask thickness it also has a maximum, which occurs at higher thickness because near-field artifacts are not included.

The optimal thickness seems to be about 0.6 mm, but this value depends on the activity of the point source for which these curves were calculated, a problem overlooked in the previous definition of the figure of merit, when activity and exposure times were fixed to a reasonable value (10 μCi and 5min) but dependence of the result on the hypothesis was not investigated. If the activity level is changed, the set of curves of Figure 7.8b is obtained. Agreement between the two FoM_t curves is good for all activity levels. This is important because it makes simulations of different activity levels redundant. Only one set of simulations, for the case of no noise, is necessary. A number of observations can be made:

FoM_p curves simply translate vertically for different activity or exposure time.

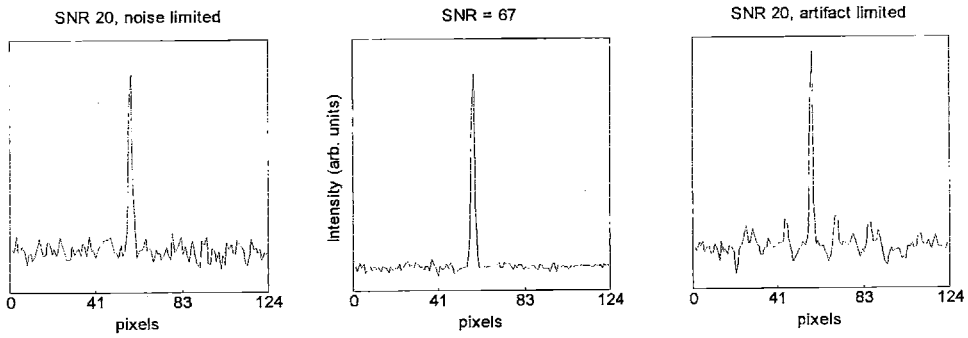


Figure 7.9: three sample PSFs for the case of the 1.66 mm resolution mask and ^{99m}Tc. From left to right, the mask thickness (transmission) is: 0.16 mm (0.60), 1 mm (0.45) and 0.35 mm (2.6×10^{-5}). Total decays for 800 s and 0.1 μCi: 0.1 (see caption of Figure 7.8). The transition from Poisson-limited SNR, to optimal SNR and artifact-limited SNR is obvious from the structure of the sidelobes and their variance.

- For a given exposure time and activity, increasing thickness leads to better FoM_t until a maximum is reached. The maximum is reached later for lower exposure times, which have poor statistics, so that better statistics due to lower transmission is worth more than increased artifacts. After the maximum the common limit FoM_a is reached by all curves. Since FoM_a is set by artifacts, it is independent of statistics.
- For any thickness, FoM_t increases with exposure time (or activity) because statistics improves, unless the artifact limit is reached.
- In most cases, a given SNR can be obtained with two masks of different thickness. The thicker one is much closer to the thickness artifact limit: longer exposure times bring little benefit. On the other hand, the thinner mask has about the same SNR, but the limit would now be set by statistics. Recalling the assumption that artifacts and noise look the same, the overall image would look the same, but in this case longer exposure times can still provide some improvement. This latter choice is, thus, preferable. A pictorial example is in Figure 7.9.

The absolute value of FoM_t is also important. Indeed one of the reasons that suggested to define it in terms of a SNR is that SNR actual values of an image can be at least intuitively understood. A reasonable image is obtained for SNR~5 while SNR~50 is an image where noise would be barely visible (Figure 7.10). In a real case a good compromise is probably a SNR of about 10. The lowest exposure time considered here is, then, already acceptable.

All curves rely on the assumption of a given mask and object-to-detector distance, in this case 40 cm. Of course, the result depends on the size of the mask hole. The rest of the study can focus on adjusting these parameters. FoM curves for the prototype mask of Chapter 6 were compared to similar curves for the same mask at a different energy (^{111}In, 245 keV and 171 keV photons) and for the high-

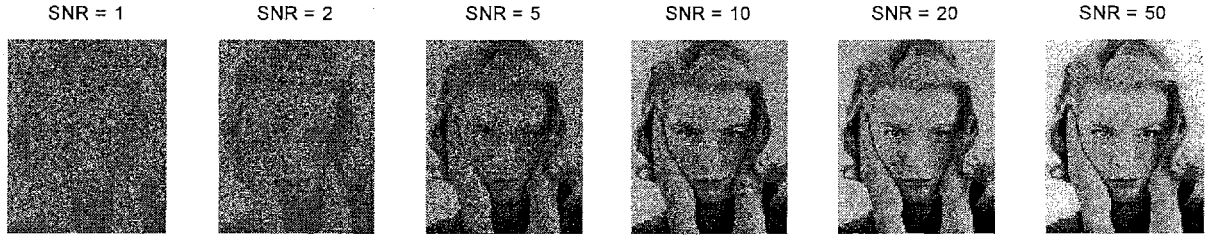


Figure 7.10: pictorial representation of the meaning of different SNR levels. Poisson statistics assumed at each point.

resolution mask of §7.1.4 at the same energy (140 keV, ^{99m}Tc) (Figure 7.11). In the comparison to a different energy, interestingly enough, the line defining the thickness artifact limit does not shift but the maximum reached for any given activity shifts to the right. One can calculate by tracing a vertical line corresponding to the thickness of the prototype mask (1.5 mm) that the maximum SNR achievable is, in both cases, about 100. However, while this limit is reached for ^{99m}Tc for 26.9 million source decays, for ^{111}In it is reached at 269 million source decays. The line also intersects other curves on the right of the maximums for ^{99m}Tc and on their left for ^{111}In. In conclusion, for the mask we built and ^{111}In, a lower SNR is expected for the same activity, the limit being set by statistics rather than by thickness artifacts. A second observation is that the mask we built may be too thick for ^{99m}Tc.

For the high-resolution mask the same SNR limit due to artifacts is reached at lower thickness, and the SNR can not be as high for the same thickness. As an example, an SNR of 60 allowed a maximum mask thickness of 2 mm for the prototype mask and of 0.8 mm for the high-resolution mask.

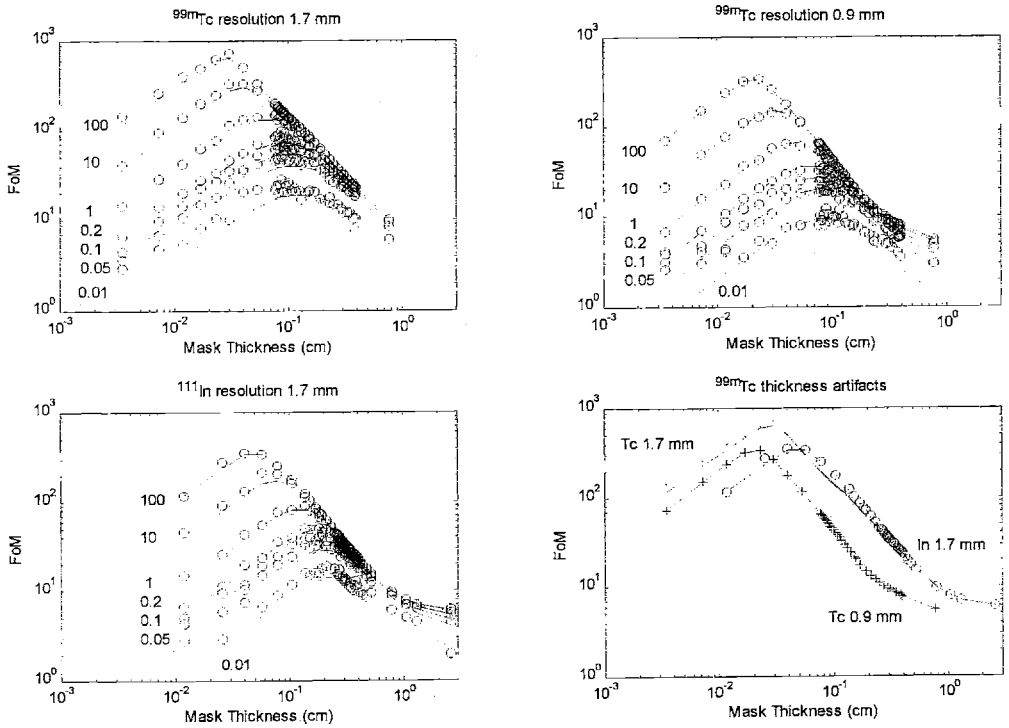


Figure 7.11: plots of the FoM_t curves (experimental and theoretical) for different levels of activity. Top left: prototype mask. Top right: mask suggested in §7.1.4. Bottom left: mask of Chapter 6 for ^{111}In, an isotope emitting γ-rays of energy higher than ^{99m}Tc (245 keV vs. 140 keV). Bottom right: direct comparison of the maximum activity curve in the three cases (2.69 billion source decays).

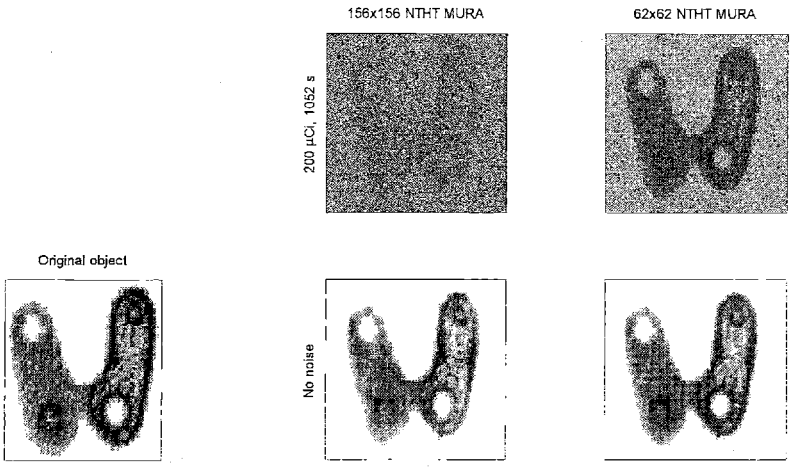


Figure 7.12: b) simulation of the thyroid phantom of a) for the high-resolution 158 × 158 mask and that designed in Chapter 6. While resolution is better, the SNR of the new mask is noticeably inferior. Activity: 200 μCi. Exposure time: 1052 s.

Consequently, the limit can be reached for 5.38 million source decays for the prototype mask but requires 26.9 million source decays for the high-resolution mask. This means that high resolution is indeed achievable, but requires a longer exposure time. An alternative point of view is that reducing the mask hole size to increase resolution shifts towards the bottom the limiting curve FoM_a: for a given thickness it may become impossible to achieve an SNR achievable with a coarser mask, no matter what the activity. The best thickness seemed to be about 0.8 mm, allowing a 9% transmission. For the same activity and exposure time, the SNR is inferior to that of the prototype mask. The result of a simulation for the same activity and exposure time is shown in Figure 7.12, where a noiseless simulation is also provided to show that better resolution is achieved if time is not an issue.

These results apply to a given object-to-detector distance only, in this case 40 cm. Simulations should be repeated for different distances. The wealth of information obtainable from these curves is such that it is difficult to cover all cases. Their use depends on the problem at hand. For instance, to overcome the limitation that they apply to point sources only, from knowledge of the object and the SNR formulae of Table 4.1, one can calculate from the SNR of a point source the SNR of an extended object. Accordingly, general conclusions are not drawn in this Chapter, which is content with having developed the general tools capable of answering a broad range of particular questions.

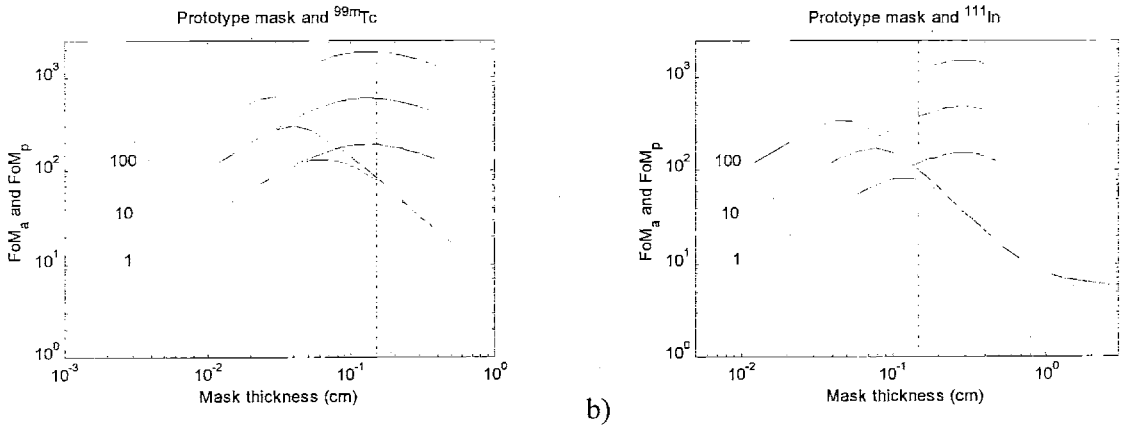


Figure 7.13: selected FoM_p and FoM_a curves from Figure 7.11. The dashed line indicates the thickness of the prototype mask.

7.5 Experimental results: ^{111}In

The curves of Figure 7.11 can predict if a mask is artifact or Poisson noise limited. Figure 7.13 predicts that a 6.6 μCi ^{99m}Tc point source imaged with the prototype mask for 735 s, which is equivalent to 179 million source decays, is Poisson limited. This is seen by looking at the value of FoM_a and FoM_p pair with parameter $179 / 29.6 \cong 6$ at the mask thickness. Since for the pair with parameter 1 FoM_p is already much higher than FoM_a, one expects the reconstruction to be artifact limited. A 36 μCi ^{111}In point source imaged with the prototype mask for 75 s, which is equivalent to a parameter of about 1 after the

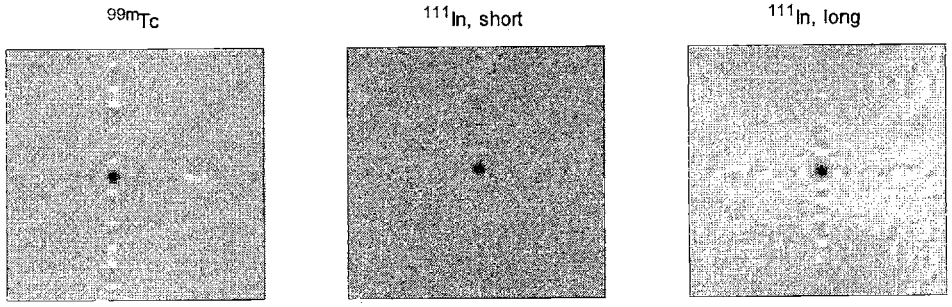
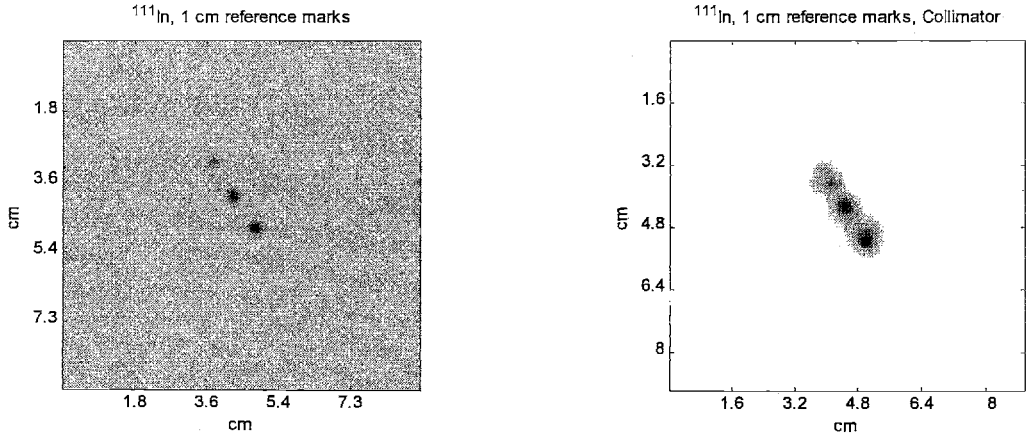


Figure 7.14: reconstruction for the ^{99m}Tc (a) and the two ^{111}In sources (short (b) and long (c) exposure) showing Poisson noise vs. thickness artifact limited reconstructions. Contrast greatly enhanced to show artifacts. In the ^{99m}Tc image the ratio between the height of the main peak and the highest side peak is 13.

lower efficiency for (^{111}In) of the detector is factored in the calculation (graphs assume 100% efficiency), is in a borderline situation were artifacts and Poisson noise are equivalent. Finally, a 30 μCi ^{111}In point source imaged with the prototype mask for 12 h (parameter ~500), should show thickness artifacts again.

Agreement with these predictions can be verified in Figure 7.14, where contrast was greatly enhanced to show artifacts. It is worth noting that in the ^{99m}Tc image the ratio between the height of the main peak and the highest side peak is 13 vs. 12.5 for the long ^{111}In image. This is in agreement with the overlap of the curves of the artifact limit at the bottom right of Figure 7.11, which predict equivalence for the prototype mask at the two energies.

A last check was that the transition to higher energy did not affect resolution. In Figure 7.15a is shown the reconstruction of three points spaced by 1 cm. The largest spot has a diameter of 2.7 mm (as measured on the image). All points are resolved and minimal artifacts appear (below the lowest point) on a uniform background. The result is compared to a medium-energy collimator in Figure 7.15b, where lower resolution is evident. This is due to the characteristics of this collimator, whose septa must be made thicker to avoid penetration, which, not only results in a resolution loss but causes artifacts as well. Figure 7.16 shows that an ^{111}In source penetrates a low-energy collimator (i.e. one designed for ^{99m}Tc) in a non-isotropic manner related to the arrangement of the collimator holes. Thicker septa can not be accommodated by simply reducing the hole diameter because this would reduce sensitivity. A complete redesign eventually leads to lower resolution. In Figure 7.16 collimators are also compared to the prototype coded aperture, which shows the same resolution for both low and medium energy.



a) b)

Figure 7.15: a) 3 marks spaced by 1 cm. Image taken with ^{111}In, two exposures for 1 million counts (about 10 min each). The sources were three drops of solution on a strip of paper: because of absorption it is difficult to know their exact diameter. b) Image of the same marks obtained with a medium energy collimator (hole diameter 2.07 mm, system resolution 10.2 mm at 10 cm)

7.6 Summary

This Chapter revisited the issue of optimal mask design and developed an advanced method for the determination of optimum mask thickness. The ultimate limit for resolution was recognized to be in the SNR. An optimum resolution mask was tentatively designed for imaging with ^{99m}Tc. Its characteristics are compared to those of the prototype mask in Table 7.4. Theory, simulation and experiment show good agreement.

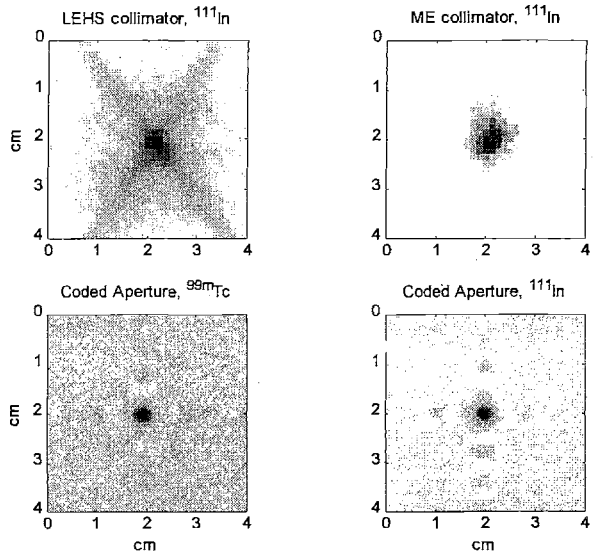


Figure 7.16: experimental results. Top left: an ^{111}In source penetrates a low-energy collimator, causing artifacts and loss of resolution. Top right: a medium energy collimator does not have artifacts but still has low resolution, due to its geometric characteristics that must accommodate thicker septa. Bottom: even though a comparison is difficult because of the different number of counts (100k for ^{99m}Tc and 267M for ^{111}In) the prototype coded aperture gives the same resolution at both energies. Also note that, due to magnification, the same field of view is covered by many more detector pixels in the case of the coded aperture (pictures were acquired with the same detector configuration in all cases).

Mask pattern		NTHT MURA 62 × 62	NTHT MURA 158 × 158
Open fraction		12.5%	12.5%
Mosaicked		yes	yes
Self-supporting		yes	yes
Mask symmetry		anti-symmetric about center	anti-symmetric about center
Material		Tungsten	Tungsten
Fabrication technology		Photo-etching	Photo-etching
Mask pixel size		1.11 mm	0.46 mm
α		2	4.082
Magnification		4.3	5.3
Resolution (at FoV = 9 cm)	Geometric	1.45 mm	0.57 mm
	System	1.66 mm	0.96 mm
Thickness		1.5 mm	0.8 mm
Attenuation at 140 keV		99%	91%
SNR	(at 592M source decays)	~20	~9
	max	~100	~45

Table 7.4: comparison of the prototype mask and the high-resolution design.

Chapter 8 EXTENSIONS AND FUTURE WORK

Future work can head in several directions. Some were explored in theory and simulation.

In section 8.1 are suggested some experiments for the mask already built that were not already done because of time constraints and limited availability of the E-Cam.

In Chapter 7, a mask capable of sub-millimeter resolution was designed but developments along lines different than simply better resolution are also possible. An example is the transition to higher energy isotopes. In section 8.2 is shown that the higher penetration encountered at higher energy may result in a considerable SNR loss. This can be recouped by prolonging exposure times, but a more practical solution is trading resolution for SNR. The example of the design of a mask for 511 keV photons is presented with a quantitative assessment of the amount of resolution that needs to be sacrificed to preserve exposure time. While this development can still rely on the ideas developed in this thesis, of which is an evolution, others require that more theoretical work be done. The most important is transition to a full 3d coded aperture system. In this Chapter is shown that only poor results can be obtained from a laminography approach, where several slices are reconstructed from data from one view. Theoretical reasons for the problem are briefly summarized in section 8.4.

8.1 Experiments with the current mask

Some obvious experiments with the prototype mask were planned but never carried out because of limited access to the E-Cam. The most interesting is perhaps a test of the dependence on the object-to-detector distance, which was always set to 40 cm in our trials on the basis of the very first optimization procedure (§6.1.4).

Moving the object closer to the detector has the promise of increasing the SNR, but artifacts (near-field and thickness) are also reinforced. Two aspects should be clarified: if 40 cm is actually the best compromise and dependence on object thickness. In fact, eq. (2.106), for a given m, becomes:

$$\frac{d\alpha}{da} = -\frac{p_m b}{p_d a^2} = -\frac{p_m}{p_d}\frac{m-1}{a} = -\frac{p_m}{p_d}\frac{(m-1)m}{z} \tag{8.1}$$

which indicates that defocusing depends on the object-to-detector distance z, the sensitivity of α on a decreases for increasing z. In other words, for high z, a thicker slice da of the object is in focus because it causes a smaller variation in $d\alpha$. In the case of small animals, this may result in better images because a considerable part of the object can be on focus, which reduces defocusing artifacts from other planes.

8.2 Extension to higher energy

At higher energy the main problem is increased mask transmission. The SNR formulae of Table 4.1 can be used to calculate the increase in exposure time needed to offset SNR losses due to penetration. Considering the prototype mask, its SNR can be rearranged to highlight the trade-off:

$$SNR_{ij}^{NTHT\ MURA} = \frac{\sqrt{NI_T/2}(1-t)\psi_{ij}}{\sqrt{(1-t)\rho + t + \xi}} \cong \sqrt{Time}\,\frac{(1-t)}{\sqrt{(1-\rho)t + \rho}} \tag{8.2}$$

where zero background was assumed. The graph of Figure 8.1 was obtained from this formula. While the time increase needed for ^{99m}Tc is only 5% higher than the time needed if the mask were perfectly opaque, it increases rapidly to 33% for ^{201}Tl and is more than 500% for ^{111}In. At 511 keV, constant SNR requires an exposure time 61 times longer. In Figure 8.3 is shown that this estimate is fairly accurate despite the

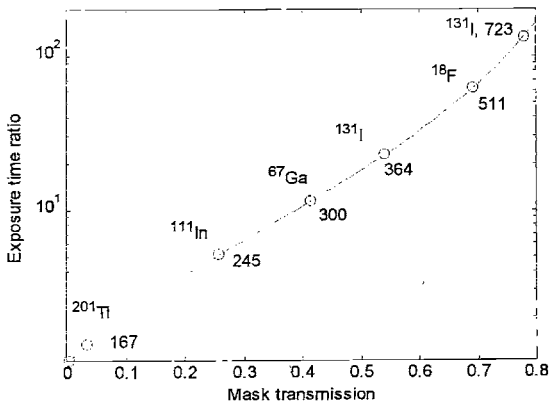


Figure 8.1: exposure time increase needed to offset a SNR loss due to mask transmission, as calculated from the analytic expression of the SNR, for a 1.5 mm-thick mask. The reference exposure time is the case of 0 transmission. The first, unlabeled, point is ^{99m}Tc (140 keV). Also shown the energy of the emitted γ rays (in keV).

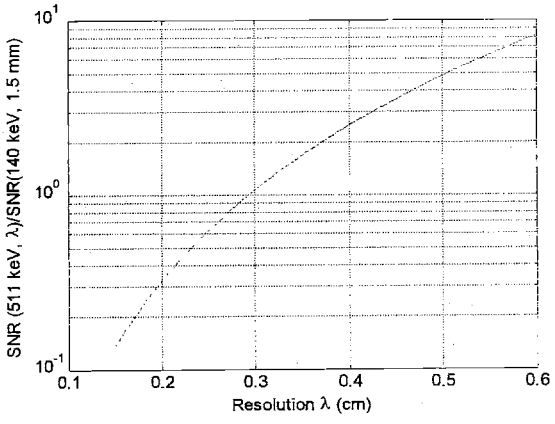


Figure 8.2: SNR at 511 keV as a function of resolution. The ratio with the SNR for a 1% (at 140 keV) transparent mask is shown. At 3 mm the SNR loss due to increased transmission is compensated.

approximations involved (mainly from mask thickness, whose effects are not included beyond the difference in transmission). The result is discouraging, even more so when the shorter half-lives of positron emitters are included in the calculations. However, the mask was not optimized for this energy. The simplest remedy is to increase mask thickness and, at the same time, use larger holes to avoid an increase of thickness artifacts, which leads to a decrease in resolution. Manipulation of the SNR formula can again give an estimate of the trade-off between resolution and SNR. The starting point is:

$$SNR_{ij}^{NTHT\ MURA} = \sqrt{\frac{NI_T}{2}} \frac{(1-t)}{\sqrt{(1-t)\rho + t}} \psi_{ij} \tag{8.3}$$

again in the assumption of no background. Only two elements of this formula depend on resolution: the transmission t and the concentration parameter ψ_{ij}. In fact, magnification is constant because the field of view and detector dimensions are constant (eq. (7.22)). The mask size also remains constant, because its projection still has to cover the whole detector. The number of open holes must change because the size of the holes is changing. However, the product NI_T equals $\rho N_T I_T$, where $N_T I_T$ is the number of photons incident on the whole mask area, which is constant, and ρ is also constant because the array family is the same, so the photon throughput does not change (first factor of eq. (8.3)).

Since m is constant, geometric resolution (the only resolution that actually matters for resolutions worse than that of the prototype mask) is proportional to the mask pixel size p_m. In first approximation

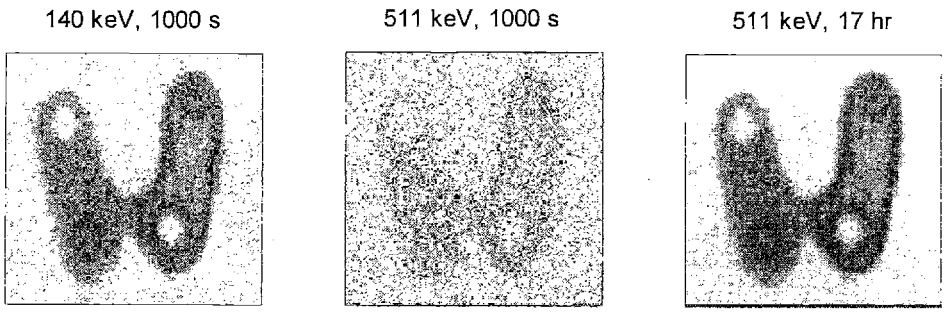


Figure 8.3: if the prototype mask is used with 511 keV photons, the SNR of the image is greatly decreased, all other things being equal, because of mask penetration. The SNR can be recouped extending the exposure time. Since the first two images were taken for 1000 s, a factor of 61 (see text) leads to 17 hours. The resulting image has SNR comparable to the ^{99m}Tc image. Decay of the source not included in the calculation of the new exposure time. This would be quite relevant for positron emitters, typically short lived.

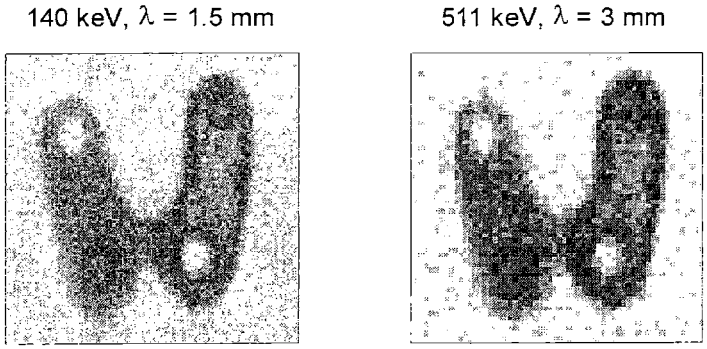


Figure 8.4: the SNR can also be recouped by reducing resolution. If holes are made larger until a resolution of 3 mm is reached, the resulting image has a SNR comparable to that obtained with ^{99m}Tc for the same time and activity. Radioactive decay of the source not included (but much less important on short exposures).

transmission can be related to p_m by assuming that the ratio of mask thickness and mask hole size is a constant. This means that also the ratio of mask thickness and resolution is a constant, which is set to the value of the prototype mask: 0.1524 / 0.1452 = 1.05. In this assumption, mask transmission is:

$$t = e^{-\mu\rho\, 1.05\, \lambda} \tag{8.4}$$

where μ is the coefficient of attenuation at 511 keV, which is 0.125 cm^2 / g. This substitution makes the second factor of eq. (8.3) a function of resolution only.

ψ_{ij} is also a function of resolution because better resolution implies that parts of the object are resolved in smaller parts of lower activity, the effect being clearly proportional to λ^2. In conclusion:

$$SNR_{ij}^{MURA\ NTHT} \propto f(\lambda)\lambda^2 \tag{8.5}$$

where $f(\lambda)$ is the second factor of eq. (8.3) where t was replaced by means of eq. (8.4). This expression can be evaluated for the generic resolution λ at 511 keV and for ^{99m}Tc for a resolution of 1.5 mm. If the ratio is taken proportionality constants are eliminated and the graph of Figure 8.2 is obtained. The equivalence point occurs at 3 mm. The calculation is again fairly accurate (Figure 8.4) despite approximations.

8.3 Artifact reduction

After near-field artifact correction, the main source of artifacts is mask thickness.

It is interesting to note that near-field artifacts are greatly enhanced by mask thickness. A second interesting observation is that, without resorting to the mask / anti-mask technique, these artifacts can be substantially attenuated by dividing the collected data by a $\cos^n(\theta)$ term with $n > 3$. In other words, a zero order correction with a higher-power cosine almost eliminates near-field artifacts. Powers as high as 8 have been found to be optimal, but the recovered image is very noisy, due to the high correction factors involved.

This observation clarified the link between near-field artifacts and mask thickness. Mask thickness alters the flux through each pinhole because of collimation of the beam (the collimation factor of section 7.4), an effect difficult to quantify but arguably related to the cosine of the incidence angle of the photons at the mask: thick masks pass preferably photons parallel to the axis of the camera and it is natural to think that this other angle-dependent attenuation will have a shape resembling that of a cosine. To attenuate the effect, pinholes may be done of varying cross-section, i.e. can be tapered to allow angled photons to still pass. In addition, nails can be inserted in the pinholes to block some fraction of photons arriving from directions parallel to the axis ([60]). The technique is actually the whole art of pinhole imaging (state-of-the-art examples are ref. [61]-[63]). All these methods do not seem to be necessary in light of the success of the anti-mask technique, but may still be worth investigating.

Artifact reduction can also be achieved with the use of non-linear method such as ART ([29], [48] and [64]) and maximum likelihood (or maximum entropy) algorithms ([23] and [65]), but these methods are computationally more demanding than simple correlation. Of course, the methods can be combined and the image produced by correlation can be used as a first iteration to accelerate the convergence of these algorithms. This may be another direction of future research, which turns out to be related to the issue of three-dimensional imaging.

8.4 Three-dimensional imaging

Coded apertures present some three-dimensional imaging potential because depth information is encoded in the size of the projection of the mask. A second way of explaining the origin of three-dimensional information is to think that each pinhole of the aperture provides a different view of the object. However, all views are restricted around the axis of the instrument. Nevertheless, this potential has been indicated as one of the major advantages of coded apertures ([13]) and a number of publications

have presented simulation and experimental results ([33], [66]-[67]). Systems relying on a single view (i.e. a single coded aperture image), however, do not seem to provide a performance as good as multiple-view systems (typically two orthogonal views ([20], [68])). Some researchers claim that maximum likelihood expectation maximization algorithms can provide improvements ([69]), but other studies report the equivalence of different reconstruction methods ([70]-[71]) and indicate that the most restrictive limit is imposed by the limited range of angles from which different views are taken, which explains the clear superiority of orthogonal-view systems.

A thorough investigation shows that the restricted view angle results in severe undersampling of some regions of the Fourier space ([72]). Criteria for successful imaging (concerning the extent of the object in real and Fourier space) have also been given starting from the formulation of a sampling theorem which shows that the finite aperture bandlimits the spatial frequencies present in the object ([73]). A more obvious limit is that 3d data are compressed in a 2d projection, which often causes a loss of information. Application of a priori knowledge, generally object outer boundary and positivity, can significantly improve the reconstructed images ([74]).

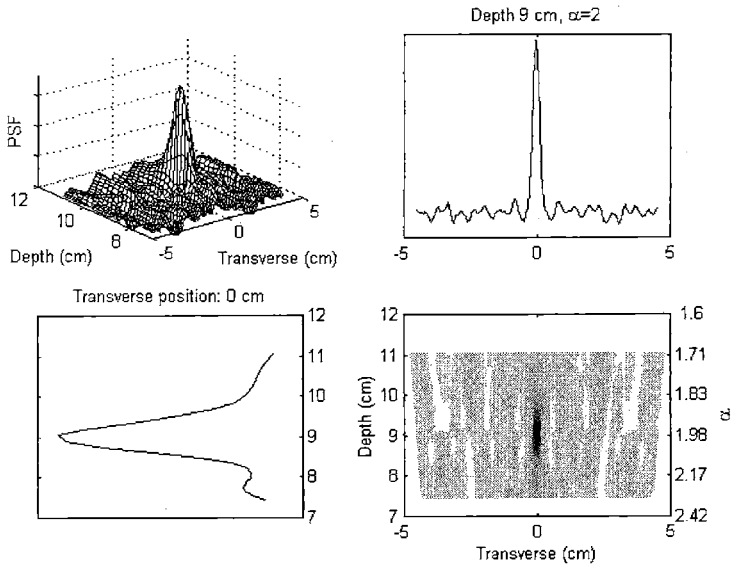


Figure 8.5: experimental three-dimensional PSF. The image at the bottom right is a color-coded top view of the 3d plot at the top left. At the other corners are two sections through the peak. The FWHM in the transverse direction is about 3 mm, due to 1.7 mm system resolution plus about 1 mm of width of the point source used. The FWHM in depth is very close to 1 cm, a factor of 5 worse than the transverse FWHM (the point source was flat, so the FWHM must be compared to 1.7 mm, not 3 mm). Note that the field of view is a function of depth. Anti-mask picture. Total number of counts 1.6×10^6 obtained in 735 s from ^{99m}Tc. Depth measured from the mask (mask-to-object distance). Mask-to-detector distance: 29.41 cm, $p_d = 2.398$ mm, $p_m = 1.114$ mm. b is different from the design value 30.7 cm because of experimental adjustments (see section 6.2).

In light of this, it should be no surprise that single-view coded aperture 3d imaging (a laminography approach, including a moving object approach) is unlikely to be successful.

An indication of this is the 3d PSF that can be obtained from any of the experiments. In Figure 8.5 is shown the reconstruction of a ^{99m}Tc source. The reconstruction is performed for different values of α, each corresponding to a different depth (distance from the mask) according to (see eq. (2.105)):

$$\alpha = \frac{p_m}{p_d}\left(1+\frac{b}{a}\right) \tag{8.6}$$

At each depth is plotted the vertical section through the reconstructed peak of the point source as a function of a transverse coordinate (a coordinate on the plane parallel to mask and detector). Note that the field of view is a function of depth. In fact, substituting in eq. (2.78) d_d, which indicates the active

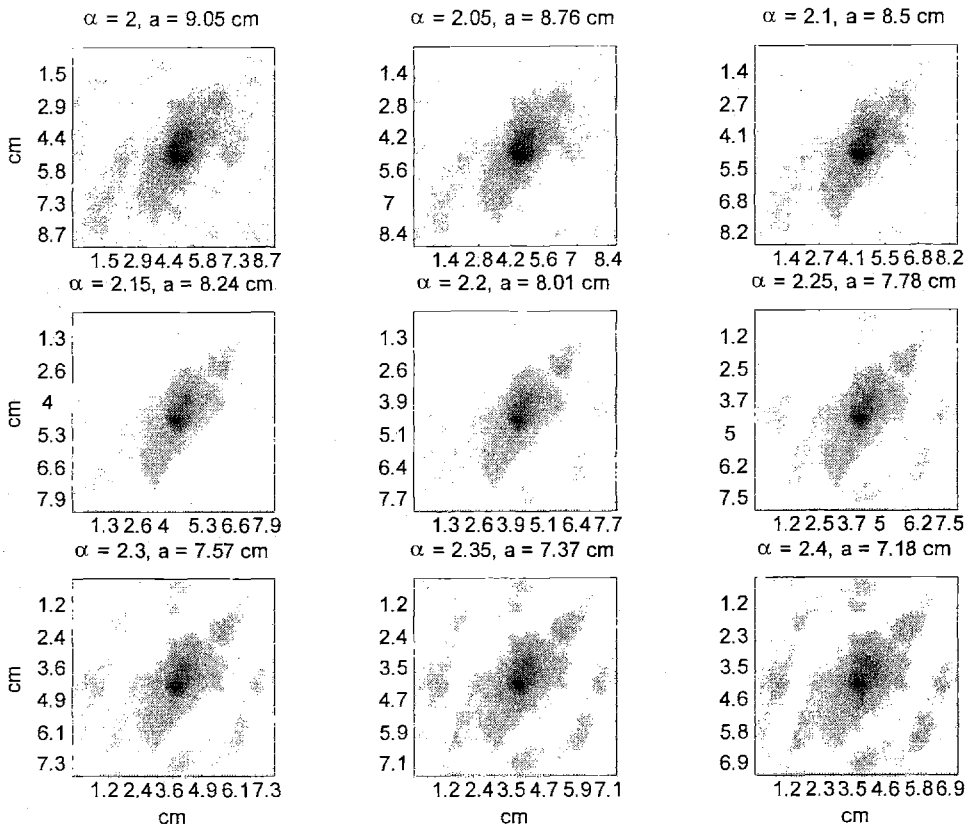


Figure 8.6: distribution of a ^{99m}Tc-labeled blood pool agent in a mouse. Decoding at different depths is an attempt at 3d reconstruction (laminography). While the image seems to go in and out of focus, the process is too smooth to crisply separate slices. Artifacts form out-of-focus planes are also most evident at the sides of the image.

area of the detector, with the size of the projection of the mask on the detector $m\,d_m$:

$$FoV = \frac{d_d}{m-1} = \frac{md_m}{m-1} = d_m\left(1+\frac{a}{b}\right) \tag{8.7}$$

where and eq. (2.74) was used to eliminate m. This shows that the field of view increases with depth, starting from the 9 cm at the design depth.

The most important result is that the FWHM in the depth dimension is a factor of 5 worse than the FWHM in the transverse plane. Also note that only for on-focus conditions (in this case $\alpha = 2$, a depth of 9 cm) the field of view is equal to the design value 9 cm and sidelobes are minimal.

A first example of 3d reconstruction was presented in section 6.3. A second is in Figure 8.6, where the data of Figure 6.12 were decoded for different depths. All slices look very much similar, due to the poor depth resolution of the system. As a result, the most visible effect of decoding at different depths is the artifacts from out-of-focus planes at the sides of the object.

It must be pointed out that in a few special cases these limited capabilities can still produce good results. This was the case of a bone scan of a rat (Figure 8.7). The rat was placed with its back facing the camera. Starting from outer planes, one can see that the knee farthest from the detector (the left one in the image) is focused first, then both knees are focused (at 10.21 cm from the mask), then the first knee is lost and starts contributing defocusing artifacts to the image (two vertical crosses centered on the two point sources of the knee). Finally, both knees are lost, but the spine is resolved in detail. Vertebrae can be seen to have a triangular shape, as it should be for young rats such as this one, a result possible only because of the high resolution of the coded aperture. Unfortunately, artifacts from out-of-plane sources corrupt the image. This corruption, however, is not severe to the point of making the image useless. Indeed, if the problem were that of locating in space an isolated point source, these artifacts may be irrelevant: almost complete 3d information is then obtained from a single view. This result is not surprising, because complete 3d information on a single point source can be obtained with a two-pinhole system, by simple ray-tracing.

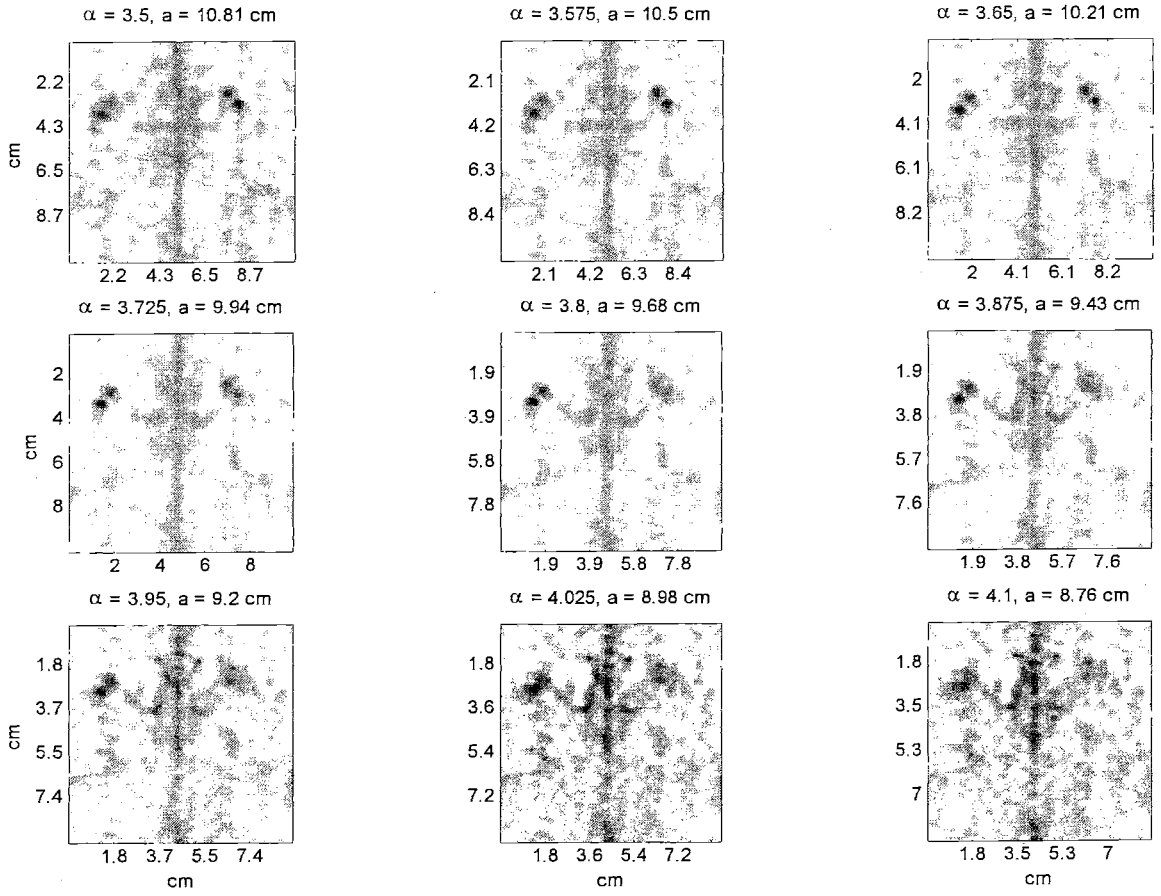


Figure 8.7: bone scan of a rat. 9 mCi injection, two 15-min images were added. Some 3d resolution is evident in the sequential focusing of the two knees (the closer to the spine is the farthest from the detector) and the spine. A collimator scan was not capable of resolving the two point sources of each knee.

Chapter 9 APPLICATION TO CAFNA

CAFNA is a bulk inspection nuclear technique based on two main technologies: coded aperture imaging and Fast Neutron Analysis. The former was discussed at length in Part I. In this Chapter FNA is introduced and some experimental results discussed. A sample CAFNA image of carbon blocks was obtained and improved with the mask / anti-mask artifact reduction method, which is again demonstrated to be useful. In the last section is designed a prototype mask for a CAFNA system, whose performance is roughly estimated.

9.1 Basic principles of CAFNA

CAFNA aims to image the spatial distribution of the elemental density of materials. It shares the major advantage of nuclear techniques of identifying materials by their elemental composition, not simply by their density as with x-ray techniques. As in FNA, in CAFNA fast neutrons are used as probing radiation. The γ-rays emitted from elements such as carbon, oxygen and nitrogen or chlorine and hydrogen after inelastic scattering or capture in the container are detected. Since the energy is peculiar to the emitting nucleus, different species (in theory isotopes) can be identified by measuring the energy of the γ-rays. The central difference between CAFNA and other neutron techniques is the method of localizing the emission point of the γ rays: a coded aperture replaces pinhole or collimator optics. The entire inspected object is simultaneously probed by a flood beam of neutrons from a commercial sealed D-T tube of the type used in the well logging industry. The tube generates neutrons using the reaction:

$$ {}_1^2\mathrm{H} + {}_1^3\mathrm{H} \rightarrow {}_2^4\mathrm{He} + \mathrm{n} \quad (\mathrm{Q} = 17.6\ \mathrm{MeV}) \tag{9.1} $$

of which about 14.6 MeV are picked up by the neutron. Typical fluxes are in the order of 10^8 neutrons per second over a solid angle of 4π and are generated with a current in the order of 100 μA at 80 kV. Since the emission location of the ensuing γ-rays (typically 1.5-11 MeV) is imaged directly through the coded aperture, neutron scattering does not affect the image and all γ-rays can be used. Furthermore, not having to tag the neutrons in space and time allows the use of the entire output of the source without compromising resolution or sensitivity. By using 14-MeV neutrons, depth penetration is high, and more

uniform penetration can be achieved by arranging multiple sources around the object. This solution would also increase flux uniformity and provide enhanced reliability through redundancy. Similarly, multiple coded aperture detectors can be used to achieve increased three-dimensional information and counting efficiency. The system is scalable in size in the sense that the number of sources and detectors can be suited to the inspection volume: CAFNA can be tailored to work with small objects, such as luggage, as well as with cargo containers.

These characteristics compare to those of another nuclear technique recently developed, Pulsed Fast Neutron Analysis (PFNA), which mechanically scans the container with a narrow beam of neutrons and determines the position of the material along the beam line by means of precision timing of the arrival of γ-rays. Since this approach assumes that the neutron interacts on the line along which it was emitted, it is potentially affected by scattering. In fact, if a neutron is scattered and then generates a γ-ray in a second scattering, the event would be placed incorrectly along the emission line of the neutron. However, in an inelastic scattering event a neutron loses energy. For example, 8 MeV neutrons do not have, after a first scattering, the energy necessary to generate inelastic scattering γ-rays in a second event, and would not present the problem. To generate neutrons of this energy the less energetic D-D reaction:

$$ {}^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{He} + \mathrm{n} \quad (\mathrm{Q} = 3.26\ \mathrm{MeV}) \tag{9.2} $$

can be used. Neutrons so generated are too slow to generate γ-rays from common elements such as carbon (4.44 MeV, which is also the threshold energy of the reaction) or oxygen (6.13 MeV). To achieve 8 MeV the incoming particle must be accelerated to several MeVs. Furthermore, short (ns) pulses must be generated because one must keep track of the departure time of the neutron to later localize the scattering event. Finally, the neutron beam must be collimated, which limits the usable fraction of particles generated. The result is that a large, laboratory-type accelerator must be used, making the technique very expensive. PFNA also requires complex mechanical motion to scan the object, and is not scalable in size because, in theory, only one neutron at a time can be present in the inspection volume.

9.2 The basics of Fast Neutron Analysis

The basis of fast neutron analysis is the inelastic reaction:

$$ {}^{\mathrm{A}}_{\mathrm{Z}}\mathrm{X} + \mathrm{n} \rightarrow {}^{\mathrm{A}}_{\mathrm{Z}}\mathrm{X} + \mathrm{n}' + \gamma \tag{9.3} $$

The threshold energy of these reactions, whose cross section is in the order of tens of millibarns (about 200 mb for the most favorable case of ^{12}C), is the energy of the emitted γ-ray, which is peculiar to the target (Figure 9.1). To identify targets, then, a spectrum of the γ rays emitted by the object is acquired. The quantity of different elements can be worked out by counting events in each full-energy peak after background subtraction but another, more efficient technique, can be used to utilize events outside the peak. The idea is to decompose the acquired spectrum in terms of some known calibration spectra.

9.2.1 Spectral deconvolution

The idea of spectral deconvolution is best explained with an example. In an exploratory experiment two sources, a ^{60}Co and a ^{137}Cs source, were used to simulate the emission of two nuclides. Four counts were taken with a $10 \times 10 \times 10\ cm^3$ NaI(Tl) scintillator connected to a PC-based Multi-Channel Analyzer via a Camac module. First, the ^{60}Co source was put on the detector and a spectrum

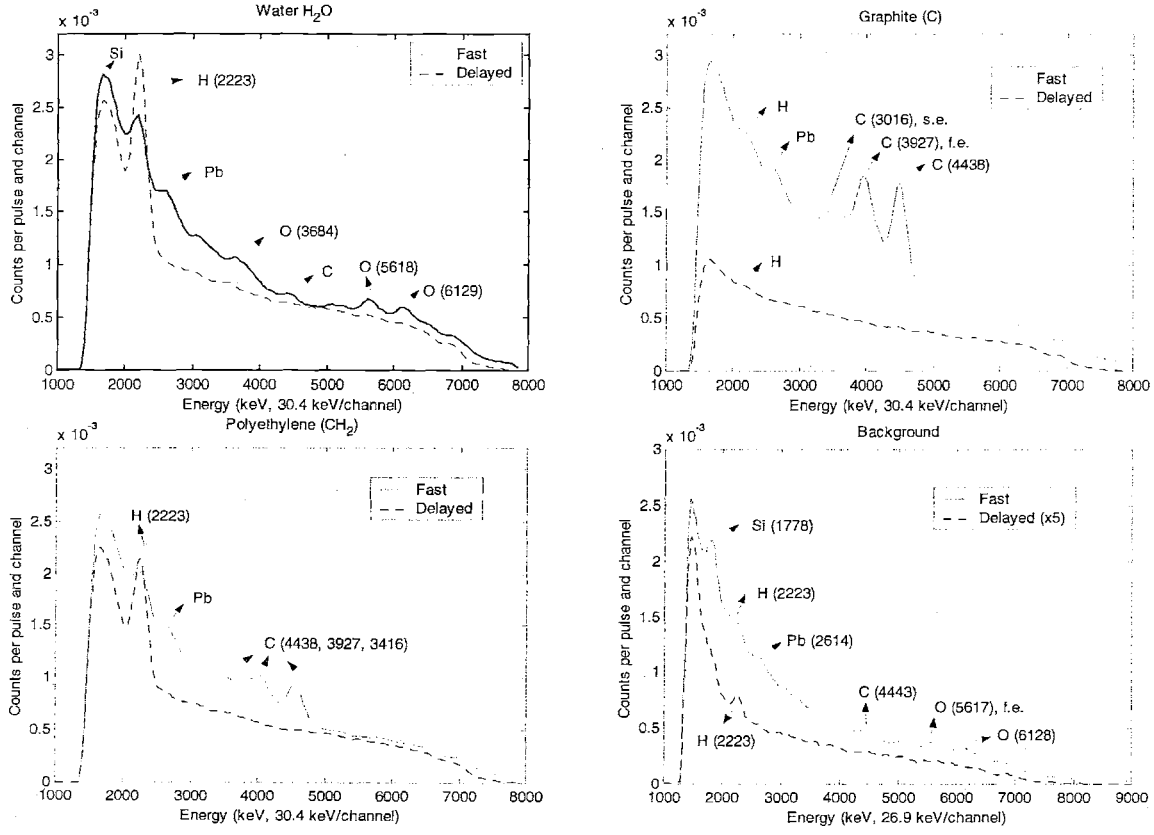


Figure 9.1: sample thermal (delayed) and fast spectra from fast neutron analysis of water, graphite and polyethylene. A background spectrum is also provided. First and second escape peaks are labeled.

acquired (Figure 9.2, top left). This was the first calibration spectrum. Then, without altering the position of the ^{60}Co source to maintain the geometry constant, the ^{137}Ce source was added on the detector and a second spectrum, the sum spectrum, was taken (Figure 9.2, top right). The third spectrum, the ^{137}Ce calibration spectrum, was collected after removing the ^{60}Co source (Figure 9.2, bottom left). Finally, all sources were removed to acquire a background spectrum. All counts were taken over a real time of 120 seconds.

The idea of spectral deconvolution is to try to reconstruct the sum spectrum from the two calibration spectra. A first guess would be that the sum spectrum is just the sum of the other two. Unfortunately, this is not true because we would be double counting background. In fact, calibration spectra are not background free, but are the sum of a spectrum due to the source alone and a background spectrum. In symbols, if naught indicates the ideal background-free spectrum:

$$Co(i) = Co_0(i) + b(i) \tag{9.4}$$

and

$$Cs(i) = Cs_0(i) + b(i) \tag{9.5}$$

where i indicates the spectrum channel. The sum spectrum, is:

$$s(i) = s_0(i) + b(i) \tag{9.6}$$

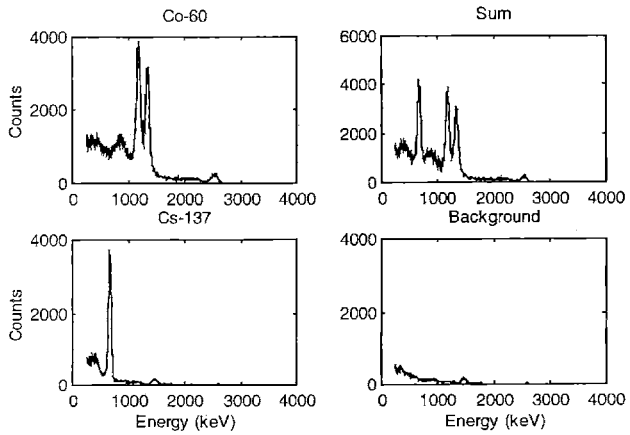


Figure 9.2: raw calibration, sum and background spectra. Calibration coefficients: slope = 5.7597 keV/channel offset = 227.8542 keV

	With no dead time correction	With dead time correction
^{60}Co	0.9864± 0.0044	1.0062 ± 0.0042
^{137}Cs	0.9013± 0.0084	0.9966 ± 0.0090
Background	-0.8913± 0.0389	-1.0065 ± 0.0398

Table 9.1: table of coefficients **k** for the example of Figure 9.2. The standard deviation was assessed with a MonteCarlo simulation that, assuming a mean value equal to the collected data, added an estimate of statistical noise and calculated the resulting variance of the coefficients.

where

$$s_0(i) = Co_0(i) + Cs_0(i) \tag{9.7}$$

while the sum of the spectra is

$$Co(i) + Cs(i) = Co_0(i) + Cs_0(i) + 2b(i) \tag{9.8}$$

which is not the sum spectrum:

$$Co(i) + Cs(i) = s_0(i) + 2b(i) \neq s(i) \tag{9.9}$$

The conclusion is that the sum is reconstructed, ideally, by summing the two calibration spectra and by subtracting background, which is making a linear combination of the three spectra Co, Cs, and b with coefficients, respectively, 1, 1, and -1. However, due to the statistical nature of nuclear counts, this is true only in the limit of a very long acquisition time or for average values. In real cases we can only verify that these coefficients are within a few standard deviations of their expected value. An estimate can be obtained by writing a system of equations with fewer variables than equations and look for the least-mean-square-error solution. In fact, for every energy channel:

$$k_1\, Co(i) + k_2\, Cs(i) + k_3\, b(i) = s(i) \tag{9.10}$$

where k_1, k_2 and k_3 are three coefficients proportional to the quantity of the corresponding element and the subscript i indicates the energy channel. In this case 499 channels were used so $i = 1, 2, \ldots, 499$ and the system is made by 499 equations in the three unknowns k_1, k_2 and k_3. The system of equations can be written in matrix form:

$$\mathbf{C}\,\mathbf{k} = \mathbf{s} \tag{9.11}$$

where $\mathbf{C}$ is a 499×3 matrix whose columns are, respectively, the cobalt, cesium and background spectra, $\mathbf{k}$ is the 3×1 column vector of components k_1, k_2 and k_3, and $\mathbf{s}$ is the 499×1 sum spectrum. The solution in the least-mean-square sense is given by (Appendix E):

$$\mathbf{k} = (\mathbf{C}^T\,\mathbf{C})^{-1}\,\mathbf{C}^T\,\mathbf{s} \tag{9.12}$$

The result for the data of Figure 9.2 is reported in the first column of Table 9.1. Even if agreement within 11% was achieved, deviation can not be explained by statistical fluctuations, because

the disagreement is in the order of 5 standard deviations or more. Indeed repeating the experiment taking four new sets of counts yielded similar results.

The problem was identified in the different activity of the two sources. Since the ^{60}Co source was stronger than the ^{137}Ce source by a factor of 3, the live count time for cobalt was lower than that of cesium, which then appears as a more intense source than it actually is. The algorithm, then, tends to compensate, reducing the relative coefficient. The dead time correction can be carried out, for all four spectra, channel by channel. For cobalt the correction formula is:

$$CoC_i = \frac{Co_i}{1 - \sum_i Co_i \frac{\tau}{T}} \tag{9.13}$$

where CoC_i is the corrected spectrum, τ is the dead time per event and T the real count time[18] (120 s in our case). For this calculation an estimate of τ is needed. It was calculated by adding two contributions, the dead time associated with a single event, which is comprehensive of detector dead time and analog-to-digital conversion time (12 μs), and the dead time associated with data transfer from the Camac memory to the PC of batch of events (32 μs = 0.0314 s per data transfer / 992 events per transfer). The total is about $\tau = 44$ μs. The corrected coefficients are in the second column of Table 9.1. The results are now within statistical deviation.

The sensitivity of the results to the value of τ is shown in Figure 9.3, where the coefficients are

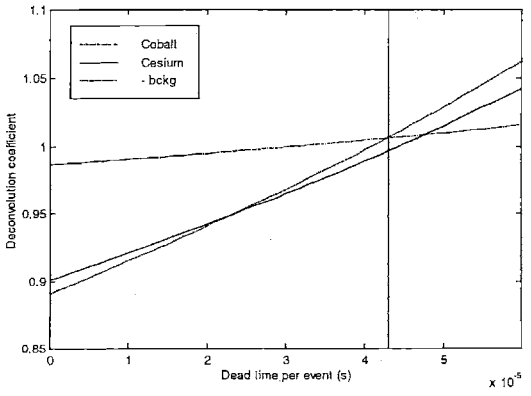


Figure 9.3: dependence of the coefficients on dead time. The value calculated for our counting system (44 μs) is shown by the vertical line.

[18] The Camac-based system did not have the option of counting for a preset live time that would have made the dead time correction unnecessary.

plotted as a function of τ. Given the accuracy of τ (a few μs), the result is fairly stable.

9.2.2 *Fast neutron analysis: experimental data*

The first task on the road to a complete FNA study was to verify the ability to generate and detect inelastic neutron scattering γ-rays. To do this, samples of different materials were exposed to fast neutrons and spectra collected in a series of 5-minute acquisitions. Graphite, water, polyethylene, sucrose and ammonium chloride were chosen because they contain all the elements present in cocaine hydrochloride, so that an appropriate mix can be used as simulant (see Table 9.2).

In the following experiments the neutron pulse repetition rate was set to 10 kHz with a duty cycle of 20%, i.e. 20 μs pulses were repeated every 100 μs, allowing an 80 μs interval between pulses. The neutron flux during the pulse was about 3×10^7 n/s. The geometry of the experiment and the pulse sequences used are in Figure 9.4.

For each material, two spectra were collected: one in correspondence of the neutron pulse ("fast" spectrum), the other with the pulse off ("thermal" spectrum). The results are in Figure 9.5, where units on the y axis are counts per pulse (and energy channel). From theory, carbon peaks are expected at 4.43 MeV, oxygen peaks at 6.13 MeV, nitrogen at 1.63, 2.31 and 5.11 MeV. Hydrogen, not having a nuclear structure, has no inelastic scattering interactions, so only capture interactions are expected. The corresponding peak is at 2.22 MeV.

By looking at the results, we can make the following comments (the uncertainty of the following experimental results is $\pm$ 1 channel = $\pm$ 30.4 keV):

Material	Chemical composition	Molecular weight	Moles per 11 moles of hydrochloride	Mole %	Weight (g)	Weight %
Graphite	C	12.011	128	64.64	1537.4	36.02
Sucrose	$C_{12}H_{22}O_{11}$	342.299	4	2.02	1369.2	32.07
Ammonium Chloride	NH_4Cl	53.491	11	5.56	588.4	13.78
Polyethylene	CH_2	14.027	55	27.78	774.0	18.13
Cocaine Hydrochloride	$C_{21}H_{22}NO_4Cl$	387.862	1	100.00	4269.0	100.00

Table 9.2: components of a cocaine hydrochloride simulant. The main goal of the choice of materials was to achieve the correct atomic composition rather than the correct mass density.

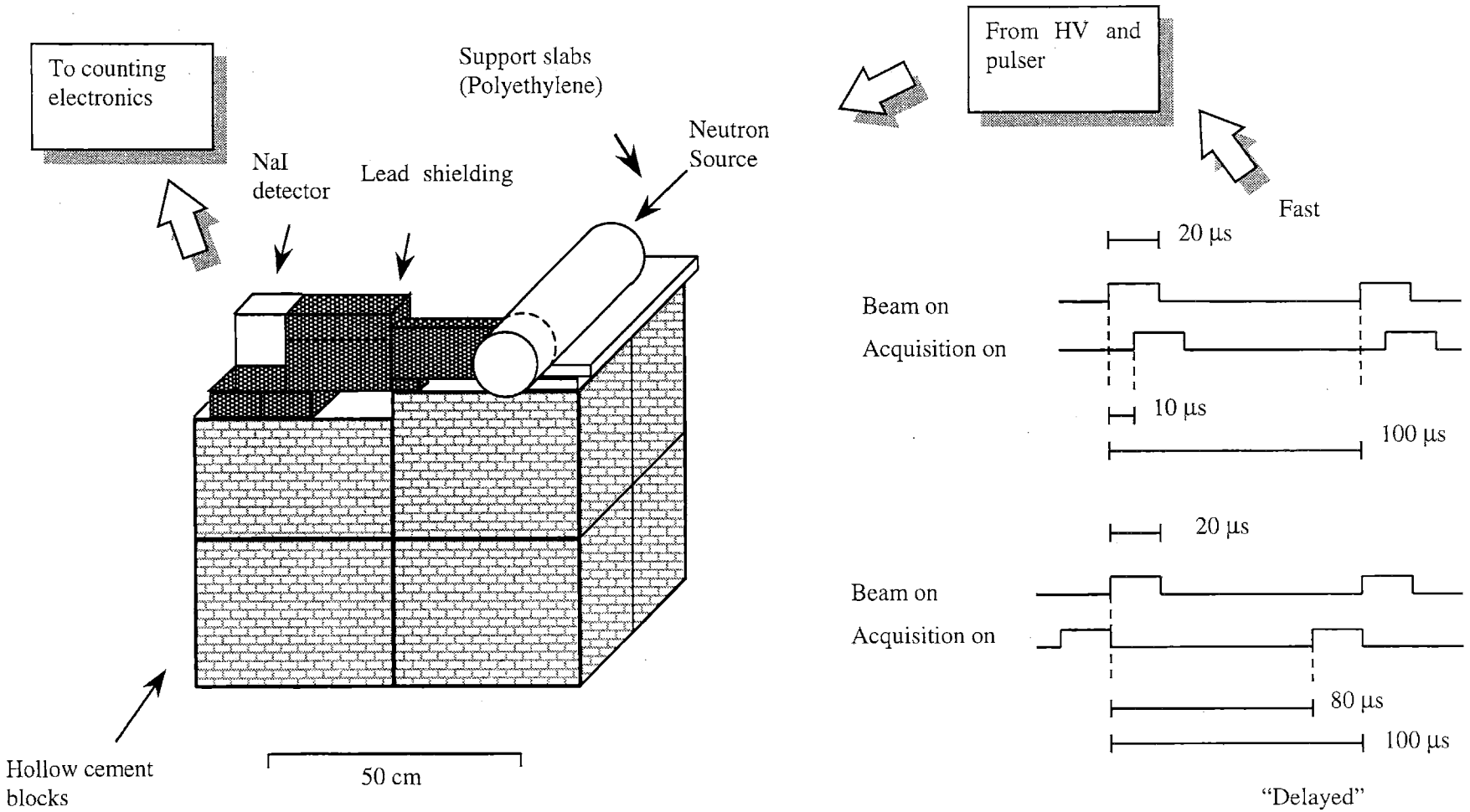


Figure 9.4: experimental setup for Fast Neutron Analysis. Different pulse sequences to acquire fast and delayed spectra are also shown. Polyethylene shielding, when used was put in place of the lead brick next to the neutron source. When variable thickness was used the tube was moved accordingly. The target is placed in front of the NaI detector.

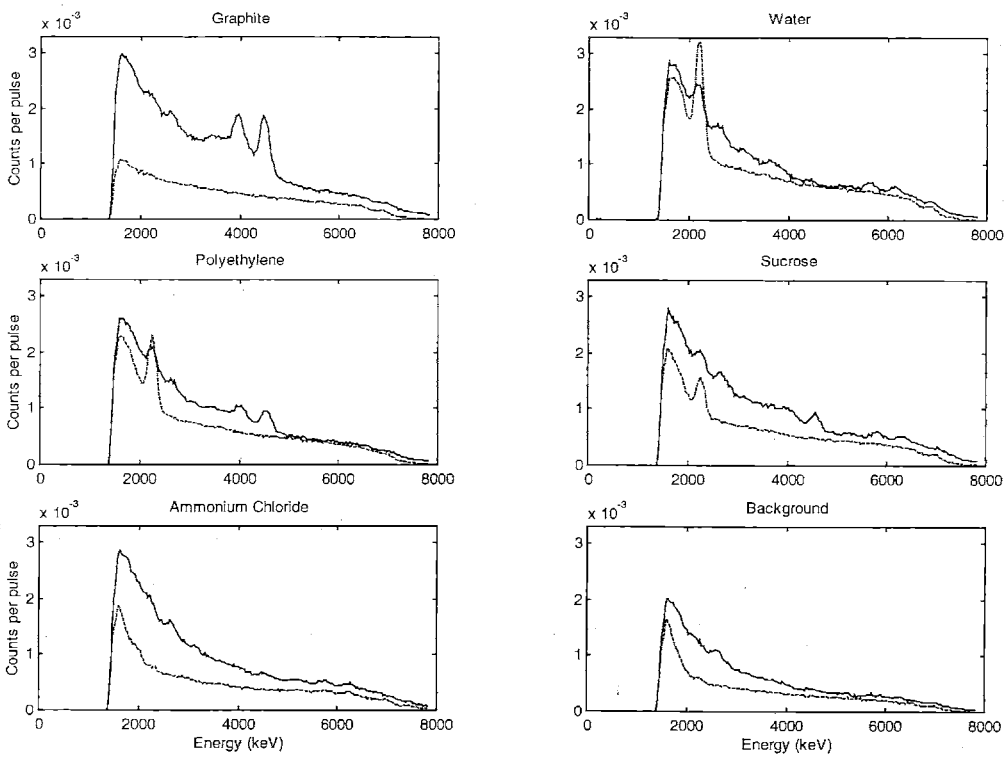


Figure 9.5: spectra collected from different materials. Continuous lines: spectrum collected with the neutron pulse on; dashed lines: spectrum collected with the neutron pulse off. 256 channels were used to collect counts. Calibration curve: E (keV) = 30.4 × channel number + 52.06.

1. Graphite shows a clear inelastic scattering peak at 4460 keV, with an escape peak at 3974 keV (expected: 4440 - 511 = 3929 keV).
2. Water shows a peak at 2210 keV in both spectra and a minor peak at 6130 keV, with its escape peak at 5675 keV (expected: 6120 - 511 = 5609 keV), due to inelastic scattering in oxygen.
3. Polyethylene shows contributions from both hydrogen (2240 keV, both spectra) and carbon (4550 keV and 4000 keV).
4. Sucrose has contributions from hydrogen (2240 keV), oxygen (6285 keV and 5800 keV) and carbon (4550 keV). Note that only the right edge of the carbon escape peak is visible. This is due to the contribution from oxygen, which fills up the space to the left of the peak, as can be seen from the water spectrum (see Figure 9.6). These counts can not be due to hydrogen, because its γ rays can not contribute above 2.22 MeV. This can

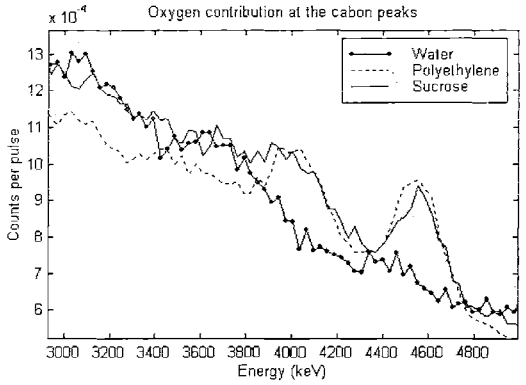


Figure 9.6: oxygen contribution in the carbon peaks region (3800 - 4440 keV). The graph is obtained by plotting on the same graph the fast spectra of water, polyethylene and sucrose.

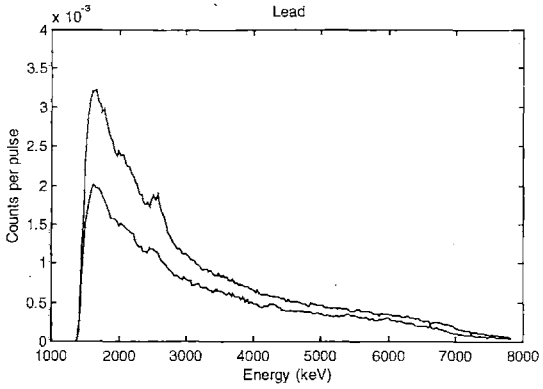


Figure 9.7: lead target (top curve) and background. The peak at 2600 keV is present in both spectra because lead shielding is always present. However it appears enhanced when a lead target is used because the contributions from target and shielding add up.

also be verified on the polyethylene spectrum, which shows two recognizable carbon peaks in spite of the presence of hydrogen.

5. Except for a peak at 2600 keV, no particular structures are recognized in the spectra for ammonium chloride. However, this peak is common to all spectra.

To confirm the suspicion that this peak may be due to the energy level 2614 keV of ^{208}Pb present in the shielding two experiments were carried out. In the first, lead was used as target material; in the

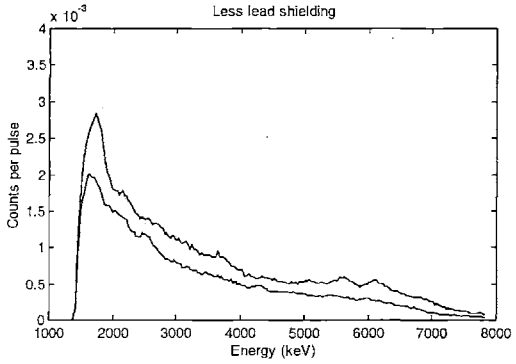


Figure 9.8: lead shielding reduction. The lower curve is a spectrum collected with all shielding in place: the detector is completely surrounded by 2 inches of lead. The higher curve was collected after removing the lead shielding not necessary to shield γ-rays from the neutron shielding placed between detector and neutron source.

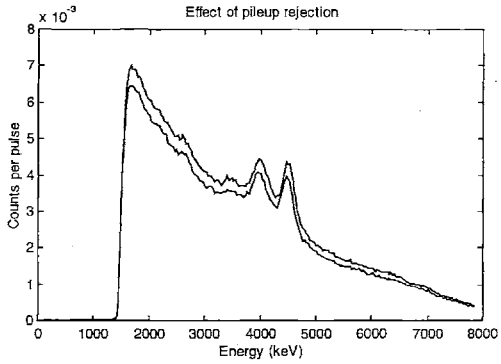


Figure 9.9: effect of pileup rejection. The pileup rejecter (lower curve) did not affect the shape of the original spectrum.

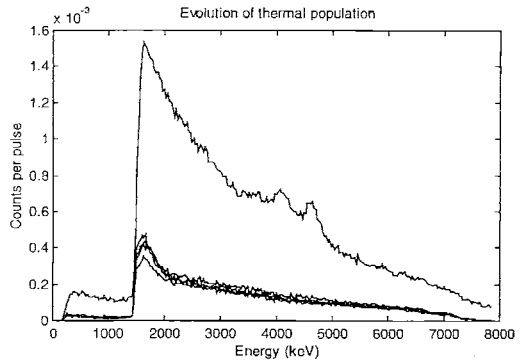


Figure 9.10: spectra acquired for different time delays of the 16 μs-gate signal: 0, 16, 32, 48 and 64 μs. Note that for the first delay the contribution of fast neutrons is still evident. Lower threshold set at 1500 keV.

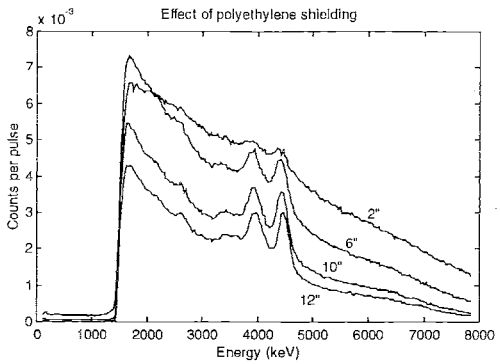


Figure 9.11: spectra collected with different thicknesses of polyethylene neutron shielding and 4 inches of lead γ shielding.

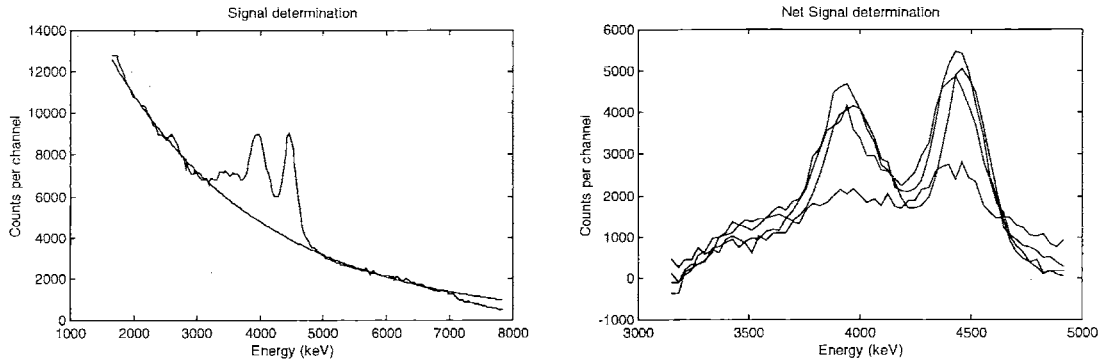
second, the amount of lead shielding, that originally surrounded the detector completely, was reduced to that necessary, on one side only, to stop the γ radiation coming directly from polyethylene neutron shielding. Using lead as a target amplifies the low energy part of the spectrum (~1200-4500 keV): in particular the peak, now at 2550 keV, is enhanced (Figure 9.7).

Reducing the amount of lead shielding around the detector confirms the result because the peaks disappear (Figure 9.8). However, as expected, background increases and other peaks come into view (2150, 5615, 6130 keV). These may be due to hydrogen and oxygen present in the room walls. This hypothesis was confirmed by moving the detector away from the wall for the following experiments.

Overall, these results confirm that the experimental setup can see γ-rays from carbon, oxygen and hydrogen. However, the neutron background is still sufficiently high to mask fast interactions on nitrogen and chlorine. Some more aspects of the system needed to be investigated to reach optimal performance.

9.2.3 System Optimization

To test if a pileup rejection circuit would be beneficial for the collection chain, two measurements of the same carbon target were taken. It can be recognized from Figure 9.9 that the count rate is sufficiently low that the only effect of the rejector is to lower the counting efficiency, without affecting the overall shape of the collected spectra. Quantitatively, an integral of 1,889,080 counts was obtained excluding the rejection circuit, versus 1,730,485 counts obtained with the rejector, for a difference of 158,595 ± 1,902 (1 σ). These numbers imply that the counts per pulse are 0.63, confirming that, at least at this neutron pulse frequency and flux, pile-up rejection is not critical.



a) b)

Figure 9.12: graphite spectrum decomposition. a) Background can be estimated by fitting and exponential curve to the data to the left and the right of the peak region. b) The curve can be subtracted from the raw data to estimate the net signal.

The dependence of the spectrum on syncronization with the neutron pulse was assessed by taking spectra after delaying a 16 μs gate by 0, 16, 32, 48, and 64 μs with respect to the shut off signal of the accelerator pulser. In Figure 9.10 the first spectrum still shows the presence of fast neutrons. This is due to the finite time needed to actually shut the accelerator beam off, which suggests not to gate the Multi-Channel Analyzer directly on the pulse signal to the accelerator but to introduce some delay. The optimal settings used for later experiments were a 20 μs gate delayed by 12 μs. All other spectra are similar in shape and magnitude, showing a very slow overall signal decay.

The effect of the thickness of polyethylene shielding was the goal of the next series of tests. Since the target is in contact with the detector, adding polyethylene is expected to reduce background counts, but also to increase the source-to-target distance, reducing signal from the target as well. Indeed the results show that fewer counts are obtained with 30.48 cm of polyethylene. However, when the thickness is reduced, the signal increases significantly but so does background. At 15.24 cm of polyethylene the background increase clearly more than offsets the signal gain (Figure 9.11). A more quantitative description can be given starting from an empirical fact. At least for graphite, peaks seem to appear on top of an exponential background that can be fitted using data from the sides of the peaks and then stripped. In Figure 9.12a the procedure is shown for one of the spectra. Signal was defined as the integral of the net (total - extrapolated background) counts in the peak region and noise was defined as the standard deviation of signal. Results are in Table 9.3 and confirm the intuition that the SNR decreases for decreasing thickness. In conclusion, 20.32 cm of polyethylene were used for the next experiments.

The dependence of the spectra on the source-to-detector distance was investigated next. As expected, the signal is strongest when the target is in contact with the detector and gradually vanishes to

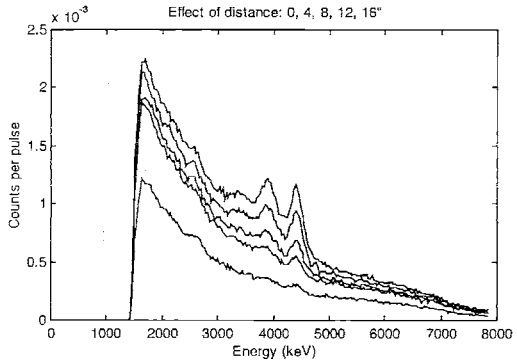


Figure 9.13: effect of target-to-detector distance on spectrum. From top to bottom, the spectra collected at 0, 5.08, 15.24, 30.48 and 40.64 cm. For increasing distance carbon peaks disappear in background.

Distance (cm)	SNR
5.08	94
15.24	140
25.4	182
30.48	179

Table 9.3: signal to noise ratio for the curves of Figure 9.12b

reach almost zero at 40.64 cm (Figure 9.13). This means that the arrangement in use, with the neutron source at the detector's side with 20.32 cm of polyethelene shielding followed by 10.16 cm of lead shielding is hardly suitable for coded aperture imaging, in light of the distance needed to accommodate the mask between object and detector.

9.2.4 Quantitative assessments

After verifying that the system was capable of recognizing the qualitative properties of the target materials, quantitative experiments were tried. The first was the simplest: two bricks of graphite were put in front of the detector in a geometry such that they would look the same to the detector as much as possible (they were on top of each other). Given the geometry, when two bricks are in front of the material, twice as many net counts were expected (Figure 9.14). The ratio of the counts can be assessed by subtracting the background spectrum and then dividing the total number of counts. The result is 1.9766 ± 0.0364, after deadtime correction.

Given the good success of the experiment above, a more challenging trial was attempted, involving more materials. Three samples of similar geometry but different compostion (water, graphite and sucrose, see Table 9.4) were irradiated for a nominal time of 10 minutes. These materials were chosen because sucrose ($C_{12}H_{22}O_{11}$) can be thought to be the sum of 12 atoms of carbon and 11 molecules of water, for a molar ratio of $12/11 \cong 1.09$. In principle, the sucrose spectrum can be reconstructed with a linear combination of water and graphite calibration spectra. The coefficients can be obtained by a least mean square technique as explained in §9.2.1.

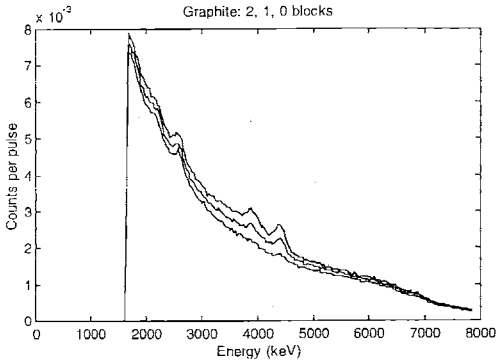


Figure 9.14: a quantitative test. The top curve shows the spectrum collected with two bricks of graphite in front of the detector. The middle curve was collected with one brick only, the bottom curve with no material at all (background count).

	Weight	mol	mol (Water)	mol (C)
Sucrose	765	2.237	24.61	26.84
Water	829	46.06	46.06	-
Carbon	1063	88.58	-	88.58

Table 9.4: weight of the samples used for the sucrose experiment

The samples were placed close to the detector. The neutron source was at the detector side, shielded by 10.16 cm of lead and 20.32 cm of polyethylene. The resulting spectra are in Figure 9.16. A comparison between the measured spectrum of sucrose and the spectrum reconstructed as linear combination of water, graphite and background is shown in Figure 9.15. Note that despite the enlarged energy scale, the reconstruction is still very good.

From the coefficients the carbon to water ratio in the sample and the total mass of the sample can be estimate. The results are in Table 9.5, where good agreement can again be verified.

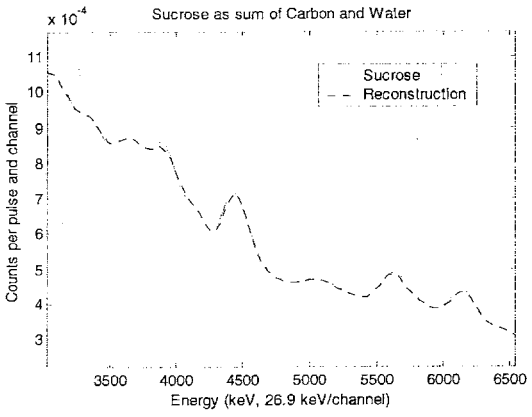


Figure 9.15: measured and reconstructed spectra of sucrose. Only the particular of the 4000 to 6500 keV zone is shown to have sufficient enlargement.

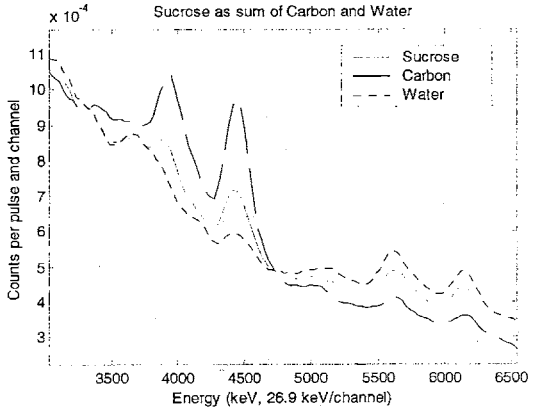


Figure 9.16: spectra collected for sucrose, carbon and water.

	Measured	Actual	Difference
Total mass (g)	864 ± 28.7	808 ± 7.09	+ 7 %
Molar ratio C / O	1.13 ± 0.07	1.09	+ 4 %

Table 9.5: total mass and molar ratio of the sucrose experiment. Comparison between expected and measured results

9.3 A 1d carbon image

The detector that will eventually be used for CAFNA is still under development. In its current configuration it is made of 64 $10 \times 10 \times 10$ cm^3 NaI(Tl) detectors arranged in a 8×8 array. All tubes are associated with a dedicated multi-channel analyzer which acquires data when a common gate signal is given. This is triggered by the sum signal over all phototubes.

To test the detector with a coded aperture a simple experiment was tried (Figure 9.17). Only a horizontal line of seven detectors facing a table was used. On the table were arranged a coded aperture and a graphite block, which, bombarded, would be the γ-ray source. The neutron generator was put immediately under the table, below the source. The detectors were shielded from direct neutrons by lead

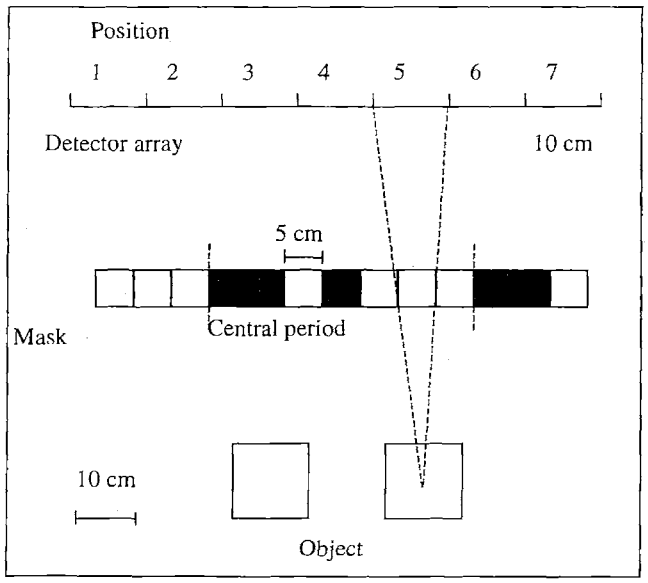


Figure 9.17: experimental setup. The neutron generator, not shown, is under the object. In the experiment with a single source, the graphite block was put exactly at the center of the object space.

blocks placed under the table, between the neutron source and the detector.

The aperture was a 1d URA of 7 positions, with four open holes. The sequence used was 0 0 1 0 1 1 1 and it was extend by 3 positions on each side to provide mosaicking. The positions were 5-cm-wide and 5-cm-thick lead was used to close opaque positions. Since the mask-to-object distance and the mask-to-detector distance were the same (25 cm), magnification was 2 (eq. (2.74)). For this crude experiment $\alpha = 1$ but sources were positioned to minimize border effects (§5.2.1). The active area of the detector was therefore 70 cm and the field of view (eq. (2.79)) 70 cm for a geometric resolution of 10 cm (eq. (2.87)). This configuration is expected to resolve the position of a $10 \times 10 \times 10$ cm block of graphite, which, not exceeding the system resolution, should appear as a point source. Two 5-minute acquisitions at 3×10^7 n/s, one without the source to assess background, were completed and the signal at each detector was defined to be the sum over all channels of the difference of the spectra acquired with and without the blocks. The signal was processed by simply taking a correlation with the decoding array, in this case coincident with the mask (Figure 9.18).

The experiment was successful because one peak is clearly visible, but sidelobes are far from being even, a deviation that can not be explained by statistical fluctuations. Since the object is a point source at the center of the field of view, zero-order correction was expected to eliminate these artifacts (§5.4.3). Decoding was repeated after dividing the signal by a $\cos^n(\theta)$ factor where θ is the incidence

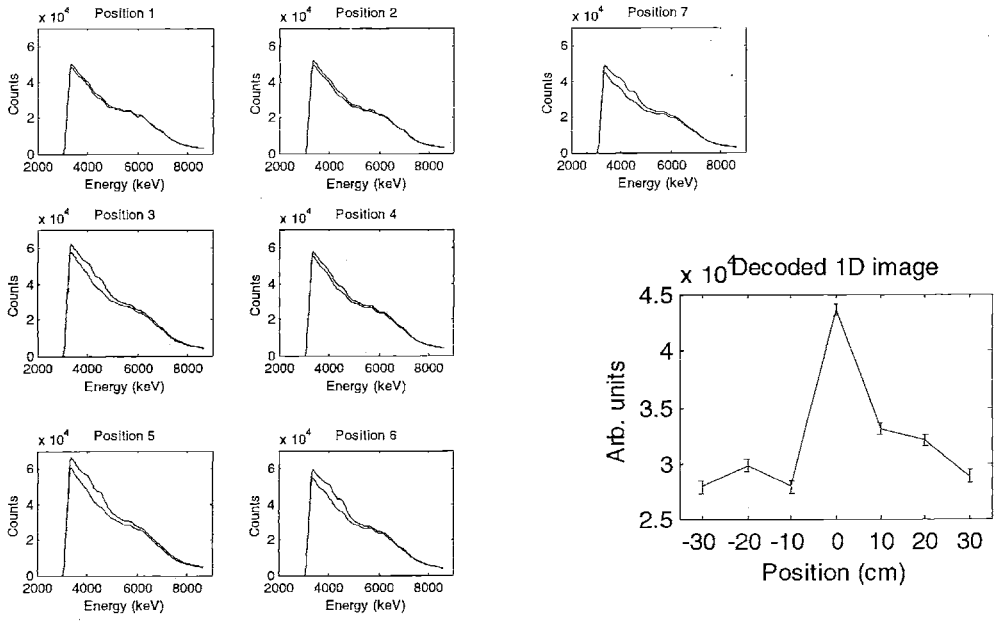


Figure 9.18: seven spectra acquired in the 1d experiment with the single graphite block. The decoded image shows uneven sidelobes with fluctuations not explained by statistical deviations.

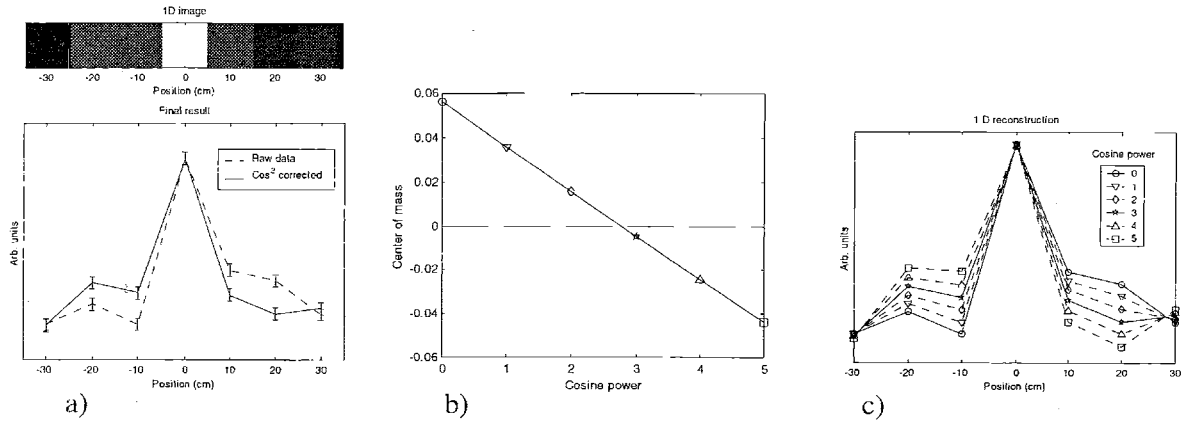


Figure 9.19: a) reconstruction of the point source of the one block experiment with and without zero order correction. b) Center of mass of the reconstructed image (ideally 0) as a function of the cosine power p. Best agreement for p = 3. c) Reconstruction for p = 0, 1, ..., 5.

angle at the center of the detector.

Figure 9.19 shows that, as expected, best results are obtained for p = 3, even though deviations are still beyond statistical fluctuation. This may be due to the poor sampling strategy ($\alpha = 1$) and mask thickness.

The next trial aimed at reconstructing the position of two sources to verify resolution. The sources were placed at 10 cm from the center of the field of view, symmetrically about the center. The two blocks were approximately the same and were illuminated in the same way by the neutron sources, that was placed under the table, below the center of the field of view (Figure 9.17). The reconstruction is at the top row of Figure 9.20. The two sources are correctly placed and resolution is confirmed. However, the relative brightness of the sources is different from expected because the symmetry of the experiment does not justify that the left peak be lower than the right beyond statistical fluctuations. Indeed repetition of the experiment showed consistently a lower left peak. Once again, however, taking an anti-mask picture and adding the results led to the expected result. This time zero-order correction is not sufficient (but still necessary) because the object is not a point at the center of the field of view.

Finally the experiment was repeated to check if the coded aperture can provide an advantage over the pinhole. From the formulae of Table 4.1 the expected SNR advantage is:

$$\frac{SNR_{URA}}{SNR_P} = \sqrt{N}\sqrt{\frac{\psi_i + \xi}{1 + 2\xi}} \tag{9.14}$$

where a half-open (M)URA pattern was assumed.

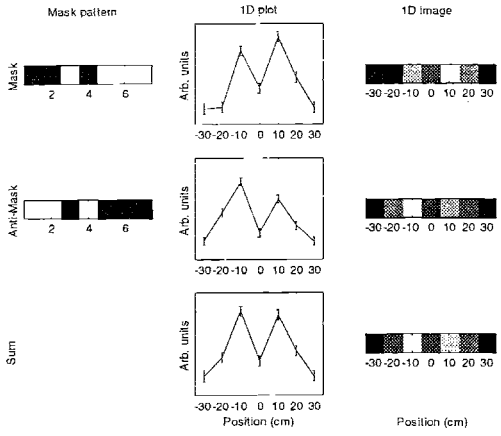


Figure 9.20: two blocks of graphite at 10 cm from the center of the field of view. Positions are reconstructed correctly but a quantitative reconstruction of intensities needs near-field artifact removal.

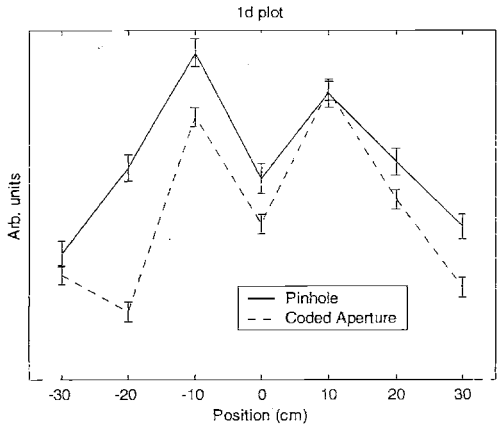


Figure 9.21: comparison of coded aperture to pinhole.

The two terms making the SNR advantage have a simple interpretation. The first is the maximum theoretical SNR advantage, obtained for $\psi = 1$ and $\xi = 0$, and is equal to $\sqrt{N}$. The second factor is always less or equal to one and represents the loss from ideality due to distributed sources and background.

In the case of two graphite blocks, $\psi = 0.5$. The parameter ξ was determined from experimental data to be about 28 and N is 4 for the mask and 3 for the anti-mask. Substitution of these values results in an SNR advantage of about 1.5, which is consistent with the approximate equivalence of the two plots of Figure 9.21. This result should come as little surprise because in this 1d example the number of open holes can not be large, so even the maximum theoretical advantage can not be large. Numbers look different in a more realistic application.

9.4 Design of a code aperture for CAFNA

A realistic problem in contraband detection could be that of finding some 50 kg of material in a cargo of size $2 \times 2 \times 2$ m^3. Assuming a density of 1.2 g / cm^3, the volume of interest is 35×35×35 cm^3: a system resolution of 20 cm is sufficient. From eq. (2.91), the mask should have 10×10 positions. Since URAs of this dimension do not exist, a 11×11 MURA pattern is chosen. It has 60 open positions, and offers a little gain in FoV at constant resolution (FoV = 2.2 m). The detector currently available is a 8×8 square of 64 $10 \times 10 \times 10$ cm^3 NaI(Tl) scintillators, which provide the necessary energy resolution.

Eq. (2.96) and eq. (2.87) lead to $p_m = 10$ cm and $m = 2$, which, from eq. (2.74), means that the mask should be midway in between object and detector. At 5 MeV, penetration is considerable (1% for 9.7 cm of lead). However, the mask pixel is 10 cm wide, and the high thickness does not lead to collimation artifacts much worse than those seen for the prototype mask.

It should be noted that the projection of the mask measures 2.2×2.2 m, but our detector is only 80×80 cm. While in a real system a larger detector can be built, we can also think of scanning the projection with our smaller detector or, better yet, move the mask so that the different parts of the projection are, in turn, shifted onto the detector. This would be more practical, because the mask, lighter and not wired, would have to be moved by smaller shifts, and advantageous, because incidence angles at the detector would be lower, reducing near-field artifacts.

Since $N = 60$, the maximum advantage is 7.75. The reduction depends on ξ and ψ. The most unfavorable case is $\xi = 0$ and uniform object, for which $\psi = 1 / N_T$ at all reconstruction points. Since the array is 50% open $\psi = 1 / 2N$: the advantage is reduced to $1/\sqrt{2}$, i.e. the pinhole is actually favored. This is actually the case in low-background applications, such as medical imaging. In a CAFNA application, however, the background due to both γ rays and neutrons is typically very high: with the experimental values $\xi = 28$ quoted above, the value of ψ, limited to 1, is irrelevant because eq. (9.14) reduces to $\sqrt{N/2}$. For a mask with 60 holes, the SNR advantage is 5.5. This value should not be underestimated because it corresponds to a 30-fold reduction in time or activity.

9.5 Summary

Good results were obtained from FNA. Some further improvements were achieved for the FNA experiment by replacing all polyethylene shielding with lead (as seen in Figure 9.4) and by smoothing the data, after background subtraction, taking the average over five channels. Sample spectra obtained with this method are in Figure 9.22. Melamine ($C_3H_6N_6$) is compared to sucrose ($C_{12}H_{22}O_{11}$). Indeed oxygen peaks are seen in the sucrose spectrum (6.128 MeV) but not in melamine and, vice-versa, nitrogen is seen in melamine (2.320 and 1.650 MeV) but not in sucrose. By comparison of the two spectra, the 2.320 MeV peak in melamine can not be due to hydrogen. This improvement, however, concerns only an experiment with the sample very close to the detector.

In Chapter 4 coded apertures were shown to have an SNR advantage over pinholes for point sources and high-background situations. While the case of medical imaging can fall under the former hypothesis, that of contraband detection may fall under the latter. In fact, neutron analysis measurements

are typically performed in a noisy environment due to the presence of the neutron source. Preliminary measurements indicate a value of the background high enough to justify use of a coded aperture. However, even though the advantage over the pinhole may still be large, the absolute value of the SNR may be very low. This may be a great difficulty to overcome especially when it is considered that in CAFNA materials are eventually identified relying on calculated ratios of elements, such as the carbon-to-oxygen ratio. To discriminate contraband from benign material, such measurements have to be within 10% or so ([2]) and propagation of uncertainty may further jeopardize the precision of the estimated ratios to the point of thwarting identification.

This discouraging picture applies to the most favorable element, carbon. When other elements are considered, it seems very dubious that the current detector will reveal them beyond 50 cm, especially when a three-dimensional object is used, with the ensuing scatter and attenuation problems.

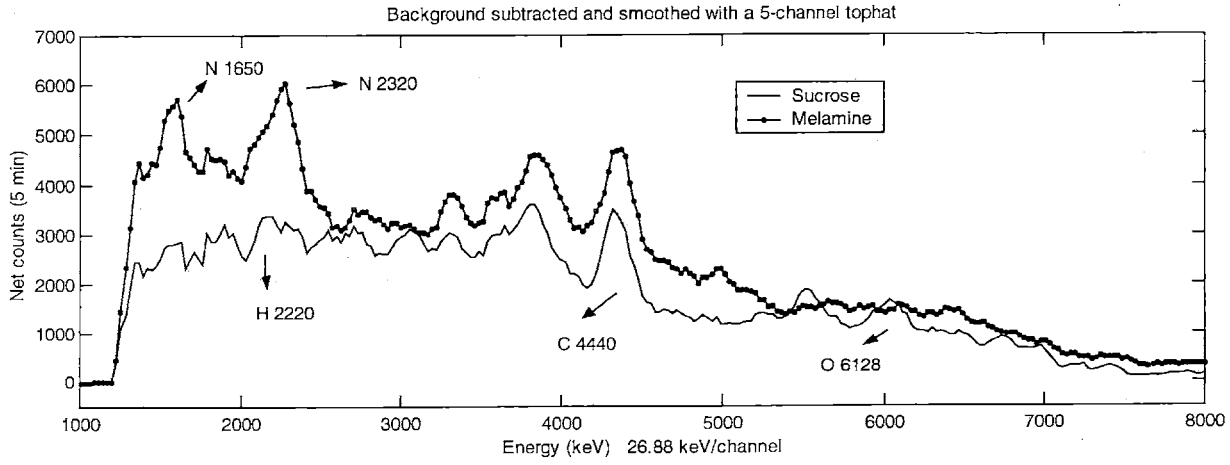


Figure 9.22: spectrum after background subtraction and smoothing over five channels. A melamine is compared to sucrose. Note the presence of nitrogen in melamine and of oxygen in sucrose. First and second escape peaks evident for oxygen and carbon. Other minor peaks can be attribute to oxygen (3.710 MeV) and nitrogen (5090 MeV).

Chapter 10 CONCLUSIONS

The tools of Nuclear Medicine available today are the arrival point of a technology largely optimized for human use. Physics has set the limits of radiation detectors in such a way that the resolution attained is sufficient for studies of adults, but poses limits on pediatric applications. This limit is even more severe in small animal imaging. Solutions proposed in literature are extreme pinhole and collimator designs that do generate beautiful images, but whose practical applications are restricted by time and activity constraints. In this thesis a sophisticated multiple-pinhole system, a coded aperture camera, was proven both theoretically and experimentally to be capable of high resolution and SNR at the same time. 1.66-mm system resolution was demonstrated and sub-millimeter resolution shown theoretically possible, under the same activity and time limits.

The result is not universal and relies on many assumptions. All were explored in detail. The first is on the field of view. Ultimately, the high resolution achieved is due to the use of a large detector in combination with a small filed of view, which allows high magnification. A second requirement for good performance is that activity in the object be concentrated in a small fraction of the field of view. Finally, the object should be thin because ideal imaging properties are limited to planar images. Despite their number, all these assumptions are satisfied in a number of practical cases. In particular we focused on mouse imaging. A mouse can easily fit in the 9×9 cm design field of view, which can be imaged with a magnification of 4.3 with a commercial Anger camera designed for human use. The thickness of a mouse is just about equal to the depth of focus of the camera, so that nearly artifact-free pictures can be taken. Typical Molecular Imaging studies are aimed at locating point sources in the body or concentrate on a part or organ of the animal and do satisfy the requirement on the spatial extent of the source. As experiments showed, the concurrence of these conditions makes mouse studies an ideal application for the coded aperture cameras designed in this thesis.

A more subtle limitation affects images where almost all activity is concentrated at a very few points. In typical cases the small activity spread around the body is still useful because it images the body, and thus provides a reference point for locating the point sources. While this works well in a collimator or pinhole study, with a coded aperture, while it is true that the bright points are imaged with high resolution and SNR, the rest of the activity is imaged with a very low SNR and may be lost in the flat statistical

noise coming from other peaks, thus making difficult the localization of the high-SNR points. However, workarounds such as targeted double injections (one to provide the shape of the body) are possible.

There were three major difficulties overcome in the design of the camera: optimization of the SNR, optimization of the dimensions of the aperture, and near-field artifact reduction. The first problem was tackled by finding a general form for results (sometimes contradictory) already available in literature and then applying it to the maximum variety of pattern families found in an extensive literature search. The experience gained in the investigation greatly helped in understanding the limitations of the SNR in coded aperture imaging. The last problem was solved within the framework of a general and original theory capable of predicting the shape and origin of artifacts in near-field coded aperture imaging. Despite the very abstract nature of the treatment, very practical remedies were proposed and were incorporated in the mask design. The mask / anti-mask technique found to attenuate near-field artifacts also turned out to be useful in attenuating mask thickness artifacts. As for the problem of optimizing the physical dimensions of the aperture, the most challenging dimension was mask thickness. A reasonably comprehensive and quantitative theory was developed and proven experimentally to be capable of predicting the factor limiting resolution, which turned out to be Poisson noise for optimally designed masks. This theory completes a logical and general design procedure for coded apertures. A much expected conclusion is that the optimal mask depends on the application at hand. For example, a second mask was designed specifically for high-resolution imaging, but we also showed how the design would have been very much different if the problem had been that of imaging a high-energy isotope. Future work on planar imaging can indeed be headed in both directions, as well as to the verification of some other assumptions of the current design (cross section of the holes, object-to-detector distance). It is finally worth mentioning the value of computer codes in the design: the prototype mask built proved to perform as predicted with remarkable agreement.

A great advantage of the coded aperture technique is that it can be retrofitted to existing technology. From an organizational point of view it is not very much different from one of the many collimators that already equip Anger cameras. This is even more valuable when it is considered that most innovative applications rely on a dedicated detector, which often entails considerable costs on top of those of the detector itself.

A coded aperture camera for planar imaging seems to be ready for application. The last steps are the construction of an automatic rotator, which must be interfaced to the data acquisition software, a flange to mount the rotator to the E-Cam, and a user-friendly software interface for the final decoding, electronic focusing and image display and data management.

The issue of three-dimensional imaging is still open for exploration. This thesis can only reach the conclusion that the approach of one-view laminography suggested in some literature is unsatisfactory. Extension to three-dimensions must pass through a multiple-view system and is fundamental to the practical use of the coded aperture technique. An application of enormous interest is, for instance, breast imaging, where small hot spots must be located with the best possible resolution and SNR for early diagnosis.

The application for which these studies were originally started, CAFNA, seems to be a very difficult case. The advantages of coded apertures are not as large as first thought, but even before any imaging is considered, the problem of extracting any signal from a large neutron background seems very serious, even for the most favorable isotopes. The problem already limits applications with isolated sources and can not but get worse when 3d objects must be located in space in the presence of self-shielding and scattering, especially in light of the accuracy needed for measurements. In any case, the detector actually available, while it seems to have reasonable spatial and energy resolution, is certainly not large enough for practical uses. Current acquisition electronics, based on Camac modules, is also a limiting factor, but little optimization has been so far attempted and is probably not the ultimate limiting factor when the more stringent physical limitations of the system are considered.

APPENDICES

Appendix A CONVOLUTION AND CORRELATION THEOREMS

This Appendix gives the proof of some theorems used in the text. Proofs are given in 1d for simplicity of notation when the extension to 2d is simple.

A.1 Definitions:

Convolution: $F * G = \int F(x)G(y-x)\,dx$

Correlation: $F \times G = \int F(x)G(x+y)\,dx$

Correlation: $F \otimes G = \int F(x)G(x \oplus y)dx$ (periodic)

$\oplus$ indicates sum modulo D. $x \oplus y$ is $x + y$ if $0 \le (x + y) < \mathrm{D}$ and the remainder of the division $(x + y)$ / D otherwise.

A.2 Theorem 1: correlation with a constant.

Be F a constant function. Then:

$$F \times G = \int F(x)G(x+y)dx = \int F \cdot G(x+y)dx = F \int G(x+y)dx = F \int G(\xi)d\xi \;\square$$

A.3 Theorem 2: $(O \times A) \otimes G = O * (A \otimes G)$

Proof:

$$(O \times A) \otimes G = \iint \left(\int_x O(x)\, A(x+y)\, dx \right) G(y \oplus z)dy = \int_x O(x) \left(\int_y A(x+y)\, G(y \oplus z)dy \right) dx$$

By replacing $\xi = x+y$

$$\int_x O(x) \int_y A(x+y)\, G(y \oplus z)dy\, dx = \int_x O(x) \int_\xi A(\xi)\, G(\xi - x \oplus z)d\xi\, dx = \int_x O(x)H(z-x)\, dx = O * H$$

where

$$H(z-x)=\int_{\xi} A(\xi)\,G(\xi-x\oplus z)d\xi$$

Since it is recognized that $H=A\otimes G$, the theorem is proven.□

A.4 Theorem 4: relationship between correlation and convolution $A\times G=\Re(A)*G=\Re(G\times A)$

Proof:

$$A\times G=\int A(x)G(x+y)dx=\begin{cases}(x\to -x) & =\int A(-x)G(y-x)dx=\Re(A)*G\\ (x+y\to z) & =\int A(z-y)G(z)dz=\int A(\eta)G(\eta+y)dz=\Re(G\times A)=A\times G\end{cases}$$

where $\Re$ indicates reflection.□

A.5 *Theorem 5:* $A\times(G*O)=(A\times G)*O$

Proof:

Since convolution is commutative:

$$A\times(G*O)=A\times(O*G)=A\times\int_x O(x)G(y-x)dx=\int_y A(y)\int_x O(x)G(z+y-x)dxdy$$

Changing the integration order:

$$A\times(G*O)=\int_x O(x)\int_y A(y)G(z-x+y)dydx=\int_x O(x)H(z-x)dx=O*H=H*O$$

where $H=A\times G$.□

A.6 *Theorem 6: cross correlation for an ideal pair* **(A, G)** *in 2d continuous representation.*

In section 2.6 we have to calculate the cross correlation:

$$A_\delta(\vec{r}-\vec{\xi})\otimes G_\delta(\vec{r})=\iint_{\vec{r}} A_\delta(\vec{r}-\vec{\xi})G_\delta(\vec{r}\oplus\vec{r}_r)d^2\vec{r} \qquad (A.1)$$

The first step is to substitute the definitions of A_δ and G_δ (eq. (2.64) and (2.82)):

$$A_\delta\otimes G_\delta=\iint_{\vec{r}}\sum_{i,j}\mathbf{A}_{i,j}\delta(\vec{r}-\vec{\xi}-\vec{r}_{i,j})\sum_{k,l}\mathbf{G}_{k,l}\delta(\vec{r}-\vec{r}_{k,l}\oplus\vec{r}_r)d^2\vec{r} \qquad (A.2)$$

Inverting summation and integration and carrying out the integration:

$$A_\delta \otimes G_\delta = \sum_{i,j} \mathbf{A}_{i,j} \sum_{k,l} \mathbf{G}_{k,l} \delta(\vec{r}_{i,j} + \vec{\xi} - \vec{r}_{k,l} \oplus \vec{r}_r) \tag{A.3}$$

It is now convenient to define: $\vec{r}_{u,v} = \vec{r}_{i,j} - \vec{r}_{k,l}$, $i = u \oplus k$ and $j = v \oplus l$. With these:

$$A_\delta \otimes G_\delta = \sum_{u \oplus k, v \oplus l} \mathbf{A}_{u \oplus k, v \oplus l} \sum_{k,l} \mathbf{G}_{k,l} \delta(\vec{r}_{u,v} + \vec{\xi} \oplus \vec{r}_r) = \sum_{u \oplus k, v \oplus l} \delta(\vec{r}_{u,v} + \vec{\xi} \oplus \vec{r}_r) \sum_{k,l} \mathbf{G}_{k,l} \mathbf{A}_{u \oplus k, v \oplus l} \tag{A.4}$$

But (**A**, **G**) is a perfect pair, so:

$$\sum_{k,l} \mathbf{G}_{k,l} \mathbf{A}_{u \oplus k, v \oplus l} = \delta(u, v) \tag{A.5}$$

where $\delta(u,v) = 1$ if $u = 0$ and $v = 0$ and 0 otherwise. Finally:

$$A_\delta \otimes G_\delta = \sum_{u \oplus k, v \oplus l} \delta(\vec{r}_{u,v} + \vec{\xi} \oplus \vec{r}_r) \delta(u, v) = \delta(\vec{\xi} \oplus \vec{r}_r) \tag{A.6}$$

Appendix B AVERAGE AND VARIANCE OF THE SIDELOBE OF A RANDOM ARRAY

B.1 Exact calculation

In §2.4.1 the distribution of sidelobe values in the auto-correlation function of a random array was shown to be:

$$p(n) = \frac{\binom{N}{n}\binom{N_T - N}{N - n}}{\binom{N_T}{N}} \tag{2.40}$$

Its average value is given by:

$$\mu_n = \sum_{n=0}^{N} n\, p(n) = \sum_{n=0}^{N} n \frac{\binom{N}{n}\binom{N_T - N}{N - n}}{\binom{N_T}{N}} = \frac{N}{\binom{N_T}{N}} \sum_{n=1}^{N} \frac{(N-1)!}{(N-n)!(n-1)!}\binom{N_T - N}{N - n} \tag{B.7}$$

Setting $m = n$-1 and $M = N$-1 we reach:

$$\mu_n = \frac{N}{\binom{N_T}{N}} \sum_{m=0}^{M} \binom{M}{m}\binom{N_T - N}{M - m} \tag{B.1}$$

Using Vandermonde's identity (see note 1 on page 40) this becomes:

$$\mu_n = \frac{N}{\binom{N_T}{N}} \binom{N_T - N + M}{M} = \frac{N^2 (N-1)!}{N_T!} \frac{(N_T - 1)!}{(N-1)!} = \frac{N^2}{N_T} = \rho N \tag{B.2}$$

where ρ is the density of the pattern N / N_T.

The variance can be reached from its definition:

$$\sigma_n^2 \equiv \sum (n-\mu_n)^2 p(n) = \sum n^2 p(n) - \mu_n^2 \tag{B.3}$$

We have to calculate:

$$\sum_{n=0}^{N} n^2 p(n) = \frac{1}{\binom{N_T}{N}} \sum_{n=0}^{N} n^2 \binom{N}{n}\binom{N_T - N}{N-n} = \frac{1}{\binom{N_T}{N}} \sum_{n=0}^{N} n \frac{N!}{(n-1)!(N-n)!}\binom{N_T - N}{N-n} \tag{B.4}$$

The substitution $m = n$-1 and $M = N$-1 comes useful again:

$$\sum_{n=0}^{N} n^2 p(n) = \frac{N}{\binom{N_T}{N}}\left[\sum_{m=1}^{M} m\binom{M}{m}\binom{N_T - N}{M-m} + \sum_{m=0}^{M}\binom{M}{m}\binom{N_T - N}{M-m}\right] \tag{B.5}$$

With the further substitutions $r = m$-1 and $R = M$-1:

$$\sum_{n=0}^{N} n^2 p(n) = \frac{N}{\binom{N_T}{N}}\left[M\sum_{r=0}^{R}\binom{R}{r}\binom{N_T - N}{R-r} + \sum_{m=0}^{M}\binom{M}{m}\binom{N_T - N}{M-m}\right] \tag{B.6}$$

and, with Vandermonde's identity:

$$\sum_{n=0}^{N} n^2 p(n) = \frac{N}{\binom{N_T}{N}}\left[(N-1)\binom{N_T - 2}{N-2} + \binom{N_T - 1}{N-1}\right] = \frac{N^2(N-1)^2}{N_T(N_T - 1)} + \frac{N^2}{N_T} \tag{B.7}$$

Substitution in eq. (B.3) and some algebra lead to:

$$\sigma_n^2 = \sum n^2 p(n) - \mu_n^2 = \frac{N^2(N_T - N)^2}{N_T^2(N_T - 1)} = \frac{N^2(1-\rho)^2}{N_T - 1} \cong \rho(1-\rho)^2 N \tag{B.8}$$

B.2 An interesting approximation

A simpler approach would be to consider that at all open positions coincidences between open holes of the array and the decoding array occur with probability ρ. The total number of coincidences would then be given by the average of a binomial process of N trials with success probability ρ. In this case the average would be ρN and the variance $\rho(1-\rho)N$. So the average is the same as in the exact calculation, but the variance is higher by a factor $(1-\rho)$. This is because in a binomial model some variance is added by the lack of the constraint that the number of open holes in the decoding array must be exactly equal to the number of holes in **A**. In other words, in the exact calculation, trials are not independent because for every success (failure) the probability of the next success decreases (increases) because a hole, i.e. some potential for future coincidences, has just been "used" ("left over"). The binomial model would be accurate if the decoding array were an entirely new random array.

Appendix C POISSON THEOREMS

C.1 Cascade of Poisson and binomial process.

In section 4.2 is stated that the cascade of a Poisson and a binomial random process is Poisson distributed with mean equal to the product of the mean of the two processes. In this specific problem, the Poisson process is one of radioactive decay and the binomial process is one of passing or not passing through the mask at an opaque position. The following proof is carried out with reference to the specific problem but its validity is general.

Proof:

The probability that x photons be emitted in a time interval T from an isotope with decay constant λ follows the Poisson distribution:

$$p(x) = \frac{e^{-\lambda T}(\lambda T)^x}{x!} \tag{C.1}$$

When the x photons hit an opaque position at the mask they can still pass through the mask with probability t (transmission coefficient). The probability that y photons pass through is conditional on the number of arrived photons x and is given by:

$$p(y \mid x) = \binom{x}{y} t^y (1-t)^{x-y} \tag{C.2}$$

The distribution of the number of photons that pass through the mask is given by the cascade of the two processes, i.e.:

$$p(y) = \sum_{x=y}^{+\infty} p(y \mid x) p(x) = \sum_{x=y}^{+\infty} \binom{x}{y} t^y (1-t)^{x-y} \frac{e^{-\lambda T}(\lambda T)^x}{x!} \tag{C.3}$$

Expansion of the binomial coefficient and rearrangement gives:

$$p(y) = \frac{t^y}{y!} e^{-\lambda T} (\lambda T)^y \sum_{x=y}^{+\infty} (1-t)^{x-y} \frac{(\lambda T)^{x-y}}{(x-y)!} \tag{C.4}$$

Substitution of $\xi = x\text{-}y$ gives:

$$p(y) = \frac{t^y}{y!} e^{-\lambda T} (\lambda T)^y \sum_{\xi=0}^{+\infty} (1-t)^{\xi} \frac{(\lambda T)^{\xi}}{\xi!} = \frac{t^y}{y!} e^{-\lambda T} (\lambda T)^y e^{(1-t)\lambda T} = \frac{(t\lambda T)^y e^{-t\lambda T}}{y!} \tag{C.5}$$

which is a Poisson distribution with mean $t\lambda T$. □

C.2 Convolution of two Poisson processes

In section 4.2 is used the result that the sum of two Poisson variables is also Poisson distributed.

Proof:

The probability that two random variables x and y, Poisson distributed with mean λ and μ add to n is given by:

$$p(n) = \sum_{x=0}^{n} p(y = n - x \mid x) p(x) = \sum_{x=y}^{+\infty} \frac{e^{-\mu} \mu^{n-x}}{(n-x)!} \frac{e^{-\lambda} \lambda^x}{x!} \tag{C.6}$$

Some rearrangement and recognition of Newton's expansion of the binomial gives:

$$p(n) = \frac{e^{-(\mu+\lambda)}}{n!} \sum_{x=0}^{n} \frac{\lambda^x \mu^{n-x} n!}{(n-x)! x!} = \frac{e^{-(\mu+\lambda)}}{n!} \sum_{x=0}^{n} \binom{n}{x} \lambda^x \mu^{n-x} = \frac{e^{-(\mu+\lambda)}}{n!} (\lambda + \mu)^n \tag{C.7}$$

which shows that n, the sum of two Poisson distributed variables is Poisson distributed with mean equal to the sum of the means. □

Appendix D SELECTED COMPUTER CODES

In this Appendix are gathered the most relevant computer codes. Codes were written for Matlab 5. These codes were meant to be fast more than readable.

D.1 Mask generation

The following program generates MURA patterns in the pattern centered shift and the associated decoding array. It also returns the symmetry of the pattern. It calls the routine `shift`.

```
% MURA2      MURA pattern generation
%       [mask,g,sym]=MURA(n) generates the p x p 2D MURA
%       returns mask, decoding array
%       and pattern symmetry (1 sym, -1 antisym)

%       s=shifted f=fast

function [mask,g,sym]=mura2sf(p)

if isprime(p)==0
   error('p must be prime')
end

ci=ismember([0:p-1],mod([1:p].^2,p))*2-1;
a=((ci'*ci)+1)/2;
a(:,1)=ones(p,1);
a(1,:)=zeros(1,p);

g=a*2-1;
g(1,1)=1;

a=shift(a,floor(p/2),floor(p/2));
g=shift(g,floor(p/2),floor(p/2));

mask=a;
sym=prod(ci([2,p]));
```

The routine `shift` is:

```
function [m]=shift(m,hs,vs);

[rm,cm]=size(m);
m=[m(rm-vs+1:rm,cm-hs+1:cm) m(rm-vs+1:rm,1:cm-hs);
   m(1:rm-vs,cm-hs+1:cm)    m(1:rm-vs,1:cm-hs)];
```

Once the mask is generated it must be mosaicked using the lines:

```
[rm,cm]=size(mask);
npm=[rm,cm];
mask=[mask(ceil(npm(1)/2)+1:npm(1),ceil(npm(2)/2)+1:npm(2)),
      mask(ceil(npm(1)/2)+1:npm(1),:),
      mask(ceil(npm(1)/2)+1:npm(1),1:floor(npm(2)/2));
   mask(:,ceil(npm(2)/2)+1:npm(2)), mask, mask(:,1:floor(npm(2)/2));
   mask(1:floor(npm(1)/2),ceil(npm(2)/2)+1:npm(2)),
      mask(1:floor(npm(1)/2),:),
      mask(1:floor(npm(1)/2),1:floor(npm(2)/2))];
```

The mosaicked mask must then be processed by the routine `cluster`, whose output should be stored in the variable `blocks`, for instance by launching `cluster` with the command: `[blocks]=cluster(mask);` . This is the clustering acceleration procedure mentioned in section 3.2.

```
function [blocks]=cluster(mask);

% Converts matrix from matrix format to rect blocks

blocks=[];
passed=zeros(size(mask));
[rm,cm]=size(mask);
for i=1:size(mask,1)
   for j=1:size(mask,2)
      if (mask(i,j)*(1-passed(i,j)))
         k=1;
         while k
            width=k;
            if (j+k<=size(mask,2));
               k=prod(mask(i,j:j+k))*(k+1)*(1-passed(i,j+k));
            else
               k=0;
            end
         end
```

```
            l=1;
            while l
               height=l;
               if (i+l<=size(mask,1));
                  l=prod(mask(i+l,j:j+width-1))*(l+1);
               else
                  l=0;
               end
            end
            yc=mean(rm/2-[i,i+height-1]+.5);
            xc=mean(-cm/2+[j,j+width-1]-.5);
            blocks=[blocks; yc xc height width];
            passed(i:i+height-1,j:j+width-1)=ones(height,width);
         end
      end
   end
```

Finally, the six variables `npm`, `rm`, `cm`, `mask`, `g` and `blocks` must be save in a file that will be loaded to import mask data in the projection routine.

D.2 Projection

The actual projection routine is launched by a larger file that prepares all inputs needed, from the mask, to all physical dimensions, the shape of the object and so forth. This file also goes on to add noise and launch decoding. The following is just an example of the (very!) many that were used. It simulates the projection of a point source through an NTHT MURA 62 × 62.

```
% Gives FoM for a point source, thick mask with several slices
% All dimensions in cm
% Mask thickness is simulated with the average fraction algorithm

clear

detzoom=1;
npix=256;
alpha=2;

Dtot=[61.4 61.4]/detzoom;
dp=Dtot/npix;
npd=min([floor([38.7 53.3]./dp);[npix npix]]);

% Fixed parameters
rd=npd(1); cd=npd(2);
```

```
D=npd.*dp;
fwhm=0.37; nbit=16;
sigma=(fwhm/(2*sqrt(2*log(2))))/dp(1);
sigma=sigma^2;

H=[9 9];

% Dependent variables

load mura62

mp=1./(npm./H+1./(alpha*dp))
m=alpha*dp./mp
d=npm.*mp
op=H./npm

npdeff=npm*alpha
Deff=dp.*npdeff

mu=1.563; rho=19.3; % At 140 keV in W (1.88 xcom)

z=40;
tr=.01;

a=z/m(1);
b=z-a;

psi=atan((Deff+H)/2/z);

% Define object

A=10; % Activity (uCi) at point (not on an area)


o=zeros(3);
o(2,2)=1;
realop=.09*[1 1];

o=o/sum(o(:))*A;
ro=size(o,1); co=size(o,2);

% Activity calculations
ster=4*atan((prod(dp)/4)/(z*norm([dp z])));
time=60; % Time in s
o=o*3.7e4*ster/(4*pi)*time;

mt=-log(tr)/mu/rho;

nslice=7;
if tr==1, nslice=1; end
```

```
mts=mt/nslice;
murhomt=mu*rho*mts;
av=linspace(a-mt/2+mt/nslice/2,a+mt/2-mt/nslice/2,nslice)

t=cputime;
pr=zeros(rd,cd);
for i=1:ro
    i/ro
   for j=1:co
      if o(i,j)

         yo=(ro/2-i+.5)*realop(1);
         xo=(-co/2+j-.5)*realop(2);

         if length(av)==1
            am=av; oa=am; b=z-am;

      prmem=pcpaffnnpotav(o(i,j),xo,yo,blocks,rd,cd,mp,mp,oa,b,D,dp,mur
      homt);
         else
            [X,Y]=meshgrid(linspace(-(D(2)-dp(2))/2,(D(2)-
      dp(2))/2,cd),linspace(-(D(1)-dp(1))/2,(D(1)-dp(1))/2,rd));
            X=X-xo; Y=flipud(Y)-yo;
            cosalpha=cos(atan(sqrt(X.^2+Y.^2)/z));

            am=av(1); oa=am; b=z-am;
            prfp=ppaffnnpotav(xo,yo,blocks,rd,cd,mp,mp,oa,b,D,dp);

            for am=av(2:length(av))
               slice=find(am==av);
               oa=am; b=z-am;

      [prfprime]=ppaffnnpotav(xo,yo,blocks,rd,cd,mp,mp,oa,b,D,dp);
               prfav=(prfprime+(prfp-
      prfprime).*exp(murhomt./cosalpha)).*(prfp<=prfprime)+(prfp+(prfpr
      ime-prfp).*exp(slice*murhomt./cosalpha)).*(prfp>prfprime);
               prfp=min(prfp,prfprime);
            end

            prmem=((1-prfav).*exp(-
      murhomt*nslice./cosalpha)+prfp).*(cosalpha.^3)*o(i,j);
         end
         pr=prmem+pr;
      end
   end
end

b=z-a;

prgeom=pr; clear prfp prfprime prfav prmem
```

```
writetime(t,'Projection time');

% Noise addition
disp('Adding noise ...')
if sum(pr(:))

   if sigma
      % Blurring
      disp(['Blurring: standard deviation (pixels) ',num2str(sigma)]),
      drawnow
      [X,Y]=meshgrid(linspace(-cd/2,cd/2,size(pr,2)),linspace(-
      rd/2,rd/2,size(pr,1)));
      gau=1/(2*pi*sigma)*fftshift(exp(-((X-(1-rem(cd,2))*.5).^2+(Y-(1-
      rem(rd,2))*.5).^2)/(2*sigma)));
      pr=cvtwo(gau,pr);
   end

   % Poisson noise
   disp(['Poisson SNR (sqrt(N)): ', num2str(sqrt(mean(pr(:))))]),
      drawnow

      pr=randp(pr.*(pr<15))+round(pr+(sqrt(pr).*randn(size(pr)))).*(pr>
      =15);

   % Dynamic range
   disp(['Dynamic range: ',num2str(nbit),' bits']), drawnow
   pr=pr+(2^nbit-1)*(pr>(2^nbit-1));

end

[X,Y]=meshgrid(linspace(-D(2)/2,D(2)/2,cd),linspace(-
     D(1)/2,D(1)/2,rd));
c=cos(atan(sqrt(X.^2+Y.^2)/z)).^3;
prc=pr./c;

% Decoding
[decim,prcut]=decd(prc,round(alpha),dp,g,1,mp,0);
```

The heart of the code is the subroutine `pcpaffnnpotav`:

```
function
     [prtot,pr,pratt]=pcpaffnnpotav(o,xo,yo,blocks,rd,cd,mp,mpt,a,b,D,
     dp,murhomt)

% p=projection, c=cosine, P=penetration, a=absolute number of counts,
     ff=very fast
% nn=No noise po=point object, t=tapered holes, av=average attenuation
```

```
% o=object (a scalar!), rd(cd)=detector rows(columns),
% define object so that its point values are uCi/prod(realop)
% mp=mask pixel size, a=object plane
% D=detector size, dp=detector pixel, N=counts per detector pixel
% mt=mask thickness, mu=attenuation coefficient

% Projection
% ----------

pr=zeros(rd,cd);

% Convert blocks from pixels in cm
blocks=[blocks(:,1)*mp(1) blocks(:,2)*mp(2) blocks(:,3)*mpt(1)
     blocks(:,4)*mpt(2)];

% Select object real position

m=1+b/a;
z=a+b;

ytm=(yo*b+D(1)/2*a)/z;
ybm=(yo*b-D(1)/2*a)/z;
xlm=(xo*b-D(2)*a/2)/z;
xrm=(xo*b+D(2)*a/2)/z;

% Sweep mask holes

for k=1:size(blocks,1)

   ytb=blocks(k,1)+blocks(k,3)/2;
   ybb=ytb-blocks(k,3);
   xlb=blocks(k,2)-blocks(k,4)/2;
   xrb=xlb+blocks(k,4);

   if ~((ytb<=ybm) | (ybb>=ytm) | (xlb>=xrm) | (xrb<=xlm))

      ytd=(ytb-yo)/a*z+yo;
      ybd=(ybb-yo)/a*z+yo;
      xld=(xlb-xo)/a*z+xo;
      xrd=(xrb-xo)/a*z+xo;

      ytdi=ceil((D(1)/2-ytd)/dp(1))+1;
      ybdi=floor((D(1)/2-ybd)/dp(1));
      xldi=ceil((xld+D(2)/2)/dp(2))+1;
      xrdi=floor((xrd+D(2)/2)/dp(2));

      top=ytdi-1-((D(1)/2-ytd))/dp(1);
      bot=(D(1)/2-ybd)/dp(1)-ybdi;
      left=xldi-1-(xld+D(2)/2)/dp(2);
```

```
      right=(xrd+D(2)/2)/dp(2)-xrdi;

      prcr=[max(1,ytdi-1):min(ybdi+1,rd)];
      prcc=[max(1,xldi-1):min(xrdi+1,cd)];

      mem=ones(ybdi-ytdi+3,xrdi-xldi+3);
      ms=size(mem);

      mem(1,:)=top*ones(1,ms(2));
      mem(:,1)=left*ones(ms(1),1);
      mem(ms(1),:)=bot*ones(1,ms(2));
      mem(:,ms(2))=right*ones(ms(1),1);
      mem(1,1)=top*left;
      mem(ms(1),1)=bot*left;
      mem(ms(1),ms(2))=bot*right;
      mem(1,ms(2))=top*right;

      % Store

      pr(prcr,prcc)=pr(prcr,prcc)+mem(prcr-ytdi+2,prcc-xldi+2);

   end
end

[X,Y]=meshgrid(linspace(-(D(2)-dp(2))/2,(D(2)-dp(2))/2,cd),linspace(-
      (D(1)-dp(1))/2,(D(1)-dp(1))/2,rd));
X=X-xo; Y=flipud(Y)-yo;
cosalpha=cos(atan(sqrt(X.^2+Y.^2)/z));
prt=(1-pr).*exp(-murhomt./cosalpha); % Transparent part
prtot=(pr+prt).*cosalpha.^3*o;
```

The subroutine `ppaffnnpotav` is different from `pcpaffnnpotav` only for the part following the last two `end` commands (a cosine correction). The routine `cvtwo` calculates a periodic convolution:

```
function [z]=cvtwo(x,y);

z=real(ifft2(fft2(x).*fft2(y)));

z=shift(z,1,1);
```

Here the routine `shift` is the same reported above. Last and least, the utility `writetime` is used to display calculation times:

```
function writetime(t,string)
```

```
%   WRITETIME    Writes execution time
%      Syntax: writetime(t,'text')
%              where t has been set earlier to cputime.

if nargin==1
   string='';
end

endtime=cputime-t;
hr=floor(endtime/3600);
min=floor((endtime-hr*3600)/60);
sec=endtime-hr*3600-min*60;
disp([string,' ',num2str(hr),'hr ',num2str(min),'m
     ',num2str(sec),'s'])
```

D.3 Decoding

Decoding is launched by the main program of D.2.

```
function [decim,pr]=decd(pr,alpha,dp,g,resc,mp,sigma)

% Decode only
% Uses delta decoding
% g enters as basic pattern (not mosaicked)

% Decoding
t=cputime;
disp('Preparing decoding array...'), drawnow

% Scale deconding matrix to size of the shadow

rex=alpha; cex=alpha;
ex=zeros(rex,cex); ex(ceil(rex/2),ceil(cex/2))=1;
dec=[];
[rm,cm]=size(g);
g=shift(g,ceil(cm/2),ceil(rm/2));
for k=1:rm
   serg=[];
   for l=1:cm
      serg=[serg g(k,l)*ex];
   end
   dec=[dec; serg];
end
clear ex k l serg
```

```
% Select center part of projection
if (((rex*rm)<size(pr,1))|((cex*cm)<size(pr,2)))
pr=pr(ceil((size(pr,1)-rex*rm)/2):ceil((size(pr,1)-rex*rm)/2)+rex*rm-
      1, ...
   ceil((size(pr,2)-cex*cm)/2):ceil((size(pr,2)-cex*cm)/2)+cex*cm-1);
end

[rd,cd]=size(pr);

% Deblurring
if sigma
disp(['Deblurring: standard deviation (pixels) ',num2str(sigma)])
[X,Y]=meshgrid(linspace(-cd/2,cd/2,size(pr,2)),linspace(-
      rd/2,rd/2,size(pr,1)));
gau=1/(2*pi*sigma)*fftshift(exp(-(X.^2+Y.^2)/(2*sigma)));
pr=ifft2(fft2(pr).*(conj(fft2(gau))./(abs(fft2(gau)))));
if max(imag(pr(:)))<1e-6
   pr=real(pr);
end
end

% Decoding
disp('Decoding ...'), drawnow
decim=ifft2(fft2(dec).*fft2(flipud(fliplr(pr))));
if max(max(imag(decim)))>1e-3*max(max(real(decim)))
   warning('Imaginary data')
end
decim=real(decim)/resc;
writetime(t,'Decoding time: '), drawnow
```

D.4 Poisson distribution

The routine mentioned in §3.2.3 to sample from a Poisson distribution considers all detector pixels independent. In each pixel is stored the mean value of the distribution to be sampled there, $\bar{r}$. First, a sample x is taken from a uniform distribution:

$$p_U(x) = \begin{cases} 1 \text{ for } 0 \le x \le 1 \\ 0 \text{ elsewhere} \end{cases} \tag{D.1}$$

then $y = -\ln(x)$ is calculated and added to z, a variable initialized to 0. The distribution of y is:

$$p_y(y) = p_U(x)\left|\frac{dx}{dy}\right| = e^{-y} \tag{D.2}$$

If $z < \bar{r}$, 1 is added to a random variable r, also initialized to 0, and another value of x sampled. Otherwise the process is stopped. In other words, r works as a counter of how many random variables distributed as y must be added before $\bar{r}$ is exceeded. r is the final result. We now have to show that it is Poisson distributed with mean $\bar{r}$.

The distribution of z after the s^{th} sampling is given by the integral of s-1 exponentials:

$$p_z^s(z) = \int_0^z \ldots \int_0^{y_3} \int_0^{y_2} e^{-y_1} e^{-y_2+y_1} e^{-y_3+y_{21}} \ldots e^{-z+y_{s-1}} dy_1 dy_2 \ldots dy_{s-1} = e^{-z} \frac{z^{s-1}}{(s-1)!} \tag{D.3}$$

The distribution of r is given by the probability that the iterations will be brought to an end after the $(r+1)^{\text{th}}$ sampling. This depends on the probability that z is some value in the interval $[z, z+dz]$ after the r^{th} sampling and the probability that the next sampled y is greater than $\bar{r}$ -z, which is:

$$p_y(y > \bar{r} - z) = \int_{\bar{r}-z}^{\infty} e^{-y} dy = e^{-\bar{r}+z} \tag{D.4}$$

Sampling of a new y does not depend on z, so probabilities can be multiplied. Since y is not limited, iterations may end at any step for any z. Integration over all $z < \bar{r}$ (otherwise the process would have already stopped) gives:

$$p_r(r) = \int_0^{\bar{r}} e^{-z} \frac{z^{r-1}}{(r-1)!} e^{-(\bar{r}-z)} dz = e^{-\bar{r}} \frac{\bar{r}^r}{r!} \tag{D.5}$$

which proves that r is Poisson-distributed with mean $\bar{r}$.

```
function r=randp(lam,m,n)

% RANDP Poisson distributed random numbers.
%    R=RANDP(LAM,M,N) generates a M-by-N matrix of Poisson distributed
%    random numbers. If LAM is a matrix, each entry of R is Poisson
%    distributed with constant LAM, according to its position.

lam=lam.*(lam>0);

if nargin==1
   [m,n]=size(lam);
end
if nargin==2
   error('Number of columns missing')
```

```
end

p=zeros(m,n);
r=zeros(m,n);

while any(any(p<lam))
    p=p-log(rand(m,n));
    r=r+1*(p<lam);
end
```

D.5 Read data from E-Cam

A number of codes were written to read and display data from the E-Cam. The E-Cam output was saved in binary format (.img) and is accompanied by a text file containing useful information (.hdr). The most general code (`hrmamcv`) imported data for both mask and anti-mask data and reconstructed the image at several depths with continuous values of α. Different rotations and reflection of the imported data had to be done depending on the particular E-Cam that acquired the data.

```
% hr   hospital reconstruction
% mam  mask antimask
% c    alpha continuous (not integer)
% v    alpha vector

% Output: several sum images

clear

% E-cam configuration

filename=input('Filename (no extension): ','s');
fid=fopen([filename '.hdr'],'r');
if fid == (-1), error(['Could not open file ' filename]); end

for l=1:36 line=fgets(fid); end % Discard first 36 lines

%detzoom=input('Zoom: ');
%npix=input('Number of pixels (scalar): ');
%dp=[61.4 61.4]/detzoom/npix;

line=fgets(fid); % Get matrix size from line 37
npix=str2num(line(20:24));

for l=38:40 line=fgets(fid); end % Discard lines 38 to 40
```

```
line=fgets(fid); % Get detectop pixel size from line 41
dp=str2num(line(33:42))/10;
dp=[dp dp];

fclose(fid); clear l line

% Calculated parameters

npd=min([floor([38.7 53.3]./dp);[npix npix]]);
cd=npd(2);
rd=npd(1);
D=npd.*dp;

% Mask selection and parameters

load mura62 g
[rm,cm]=size(g);
npm=[rm cm];

b=input('Mask-detector distance= ');

d=6.9092; %input('Elementary mask dimension (horizontal)= ');
d=[d/npm(2)*npm(1) d];
mp=d./npm;

% Imports .img files

fid=fopen([filename '.img'],'rb','ieee-be');
pri=fread(fid,[npix npix],'int16')';
%pri=fliplr(pri);
fclose(fid); clear fid

filename=input('Filename of .img file, rotated (no extension): ','s');
fid=fopen([filename '.img'],'rb','ieee-be');
prir=fread(fid,[npix npix],'int16')';
%prir=fliplr(prir);
fclose(fid); clear fid

%load test
%pri=o;
%prir=or;

s=round((size(pri)-npd)/2);
s=[s(1)+1 s(1)+npd(1) s(2)+1 s(2)+npd(2)];

pr=pri(s(1):s(2),s(3):s(4));
prr=prir(s(1):s(2),s(3):s(4));
```

```
alphav=input(['Reconstruction plane
      (',num2str(max(mp./dp)),'<=\alphav<=',num2str(min(size(pr)./npm))
      ,'): ']);
l=length(alphav);

m=(dp./mp)'*alphav;
H=(ones(length(alphav),1)*d)'./(1-1./m);
a=b./(m(1,:)-1);
z=a+b;
op=(dp(1)'.*alphav)./(m(1,:)-1);

q=ceil(sqrt(l));
r=ceil(l/q);

% Reconstruction

hs=0; vs=0;

decims=zeros([size(pr),length(alphav)]);
decimsr=zeros([size(pr),length(alphav)]);
decimsizes=[];
[x,y]=meshgrid(linspace(-(D(2)-dp(2))/2,(D(2)-
      dp(2))/2,npix),linspace(-(D(1)-dp(1))/2,(D(1)-dp(1))/2,npix));
for alpha=alphav;
   c=cos(atan(sqrt(x.^2+y.^2)/z(find(alpha==alphav)))).^3;
      c=c(s(1):s(2),s(3):s(4));
   prc=pr./c;
   prrc=prr./c;
   decim=decdcc(prc,shift(g,0,1),alpha,hs,vs);
   decimr=decdcc(prrc,rot90(g),alpha,hs,vs);
   decims(1:size(decim,1),1:size(decim,2),find(alpha==alphav))=decim;

      decimsr(1:size(decimr,1),1:size(decimr,2),find(alpha==alphav))=de
      cimr;
   decimsizes=[decimsizes size(decim)'];
end
decims=decims(1:max(decimsizes(1,:)),1:max(decimsizes(2,:)),:);
decimsr=decimsr(1:max(decimsizes(1,:)),1:max(decimsizes(2,:)),:);

% Display result

figure

for alpha=alphav;
   num=find(alpha==alphav);
   subplot(r,q,num)
   decim=decims(1:decimsizes(1,num),1:decimsizes(2,num),num);
   decimr=decimsr(1:decimsizes(1,num),1:decimsizes(2,num),num);
   im=decim+decimr;
   imagesc(im);
```

```
    caxis([0 max(im(:))])
    axis image, title(['\alpha = ',num2str(alpha),', a =
       ',num2str(round(a(num)*100)/100),' cm']), colormap hot

       set(gca,'XTicklabel',round(str2num(get(gca,'XTickLabel'))*op(num)
       /alpha*10)/10)
    if (ceil(num/q)==r), xlabel('cm'),end

       set(gca,'YTicklabel',round(str2num(get(gca,'YTickLabel'))*op(num)
       /alpha*10)/10)
    if rem(num,q)==1, ylabel('cm'), end
end

colormap hot
zoom on

clear sigma filename t rm rd cm cd decim decimr im
```

The important part of the code is the decoding subroutine decdcc:

```
function [decim,pri]=decdcc(pr,g,alpha,hs,vs)

% dec Decode only
% d   uses delta decoding
% c   alpha assumes continuous values
% c   closest integer value of alpha
% g   enters as basic pattern (not mosaicked)

if nargin==3, hs=0; vs=0; end

t=cputime;

alphai=round(alpha);

[rd,cd]=size(pr);
[rm,cm]=size(g);
g=shift(g,ceil(cm/2),ceil(rm/2));

% Scale deconding matrix to integer part of alpha

dec=zeros([rm,cm]*alphai);
dec(ceil(alphai/2):alphai:size(dec,1),ceil(alphai/2):alphai:size(dec,2
    ))=g;

% Select center part of projection

if (((alpha*rm)>size(pr,1))|((alpha*cm)>size(pr,2)))
```

```
   disp('\Projection too large!');
end

 [xi,yi]=meshgrid([floor((cd-
     alpha*cm)/2):alpha/alphai:(cd+alpha*cm)/2],...
                 [floor((rd-
     alpha*rm)/2):alpha/alphai:(rd+alpha*rm)/2]);
xi=hs+xi(1:cm*alphai,1:cm*alphai);
yi=vs+yi(1:rm*alphai,1:rm*alphai);
pri=interp2(pr,xi,yi,'*linear');

% Decoding

disp('Decoding ...')
decim=ifft2(fft2(dec).*fft2(flipud(fliplr(pri))));
if max(max(imag(decim)))>1e-3*max(max(real(decim)))
   warning('Imaginary data')
end
decim=real(decim);

writetime(t,'Decoding time: ')
```

A sample header file can help in understanding the way camera setup data are imported by `hrmamcv`. The line `!INTERFILE :=` is the actually first of the file.

```
!INTERFILE :=
!imaging modality := nucmed
!originating system := SIEMENS
!version of keys := 3.3
date of keys := 1992:01:01
conversion program := ICONInterfile
program author := Siemens Medical Systems - Nuclear Medicine Group
program version := 2.01
program date := 1994:04:22
!GENERAL DATA :=
original institution := UNKNOWN
contact person := SMS Hotline - (800)873-3582
!data offset in bytes := 0
!name of data file := image7.IMG
patient name := CODED APERATURE 1
!patient ID := 1
patient dob :=
patient sex := M
!study ID := Coded Aperature
exam type := obj -det 38.6
data compression := none
data encode := none
```

```
!GENERAL IMAGE DATA :=
!type of data := Static
!total number of images := 1
study date := 2000:07:18
study time := 18:25:07
imagedata byte order := BIGENDIAN
!number of energy windows := 1
energy window[1] := Tc99m
flood corrected := N
decay corrected := N
!STATIC STUDY (General) :=
number of images/energy window := 1
!Static Study (each frame) :=
!image number := 1
!matrix size [1] := 256
!matrix size [2] := 256
!number format := unsigned integer
!number of bytes per pixel := 2
scaling factor (mm/pixel) [1] :=    2.398
scaling factor (mm/pixel) [2] :=    2.398
image duration (sec) :=  734.622
image start time :=    0.000
maximum pixel count := 233
total counts := 1600004
!END OF INTERFILE :=
```

Appendix E LEAST-MEAN-SQUARE ERROR SOLUTION OF A LINEAR SYSTEM

The least-mean-square error solution of the linear system:

$$\mathbf{C\,k = s} \tag{E.1}$$

was stated in §9.2.1 to be:

$$\mathbf{k} = (\mathbf{C}^T\,\mathbf{C})^{-1}\,\mathbf{C}^T\,\mathbf{s} \tag{E.2}$$

Proof:

Define the square error:

$$E = (\mathbf{C\,k - s})^T\,(\mathbf{C\,k - s}) \tag{E.3}$$

which is the sum of the squares of all the reconstruction errors:

$$k_1\,Co(i) + k_2\,Cs(i) + k_3\,b(i) - s(i) \tag{E.4}$$

The error (and, thus, the mean error) is minimized if:

$$\frac{\partial E}{\partial \mathbf{k}} = \mathbf{0} \tag{E.5}$$

where $\partial E / \partial \mathbf{k}$ is the column vector of the three derivatives $\partial E / \partial k_j$. However,

$$\frac{\partial E}{\partial \mathbf{k}} = 2\,\mathbf{C}^T(\mathbf{Ck} - \mathbf{s}) = \mathbf{0} \tag{E.6}$$

implies:

$$\mathbf{C}^T\,(\mathbf{Ck} - \mathbf{s}) = \mathbf{0} \tag{E.7}$$

which leads to the solution with a little matrix algebra. In fact

$$\mathbf{C}^{\mathrm{T}}\,\mathbf{C}\,\mathbf{k} = \mathbf{C}^{\mathrm{T}}\,\mathbf{s} \tag{E.8}$$

where $\mathbf{C}^{\mathrm{T}}\,\mathbf{C}$ is square and can be inverted ($\mathbf{C}$ is the matrix of calibration spectra, which, being independent vectors, make $\det(\mathbf{C}^{\mathrm{T}}\,\mathbf{C}) \neq 0$) and multiplied to the left of both sides to get the solution. The critical point is actually a minimum because:

$$\frac{\partial^2 E}{\partial^2 \mathbf{k}} = \mathbf{C}^{\mathrm{T}}\mathbf{C} > \mathbf{0}\ \square \tag{E.9}$$

BIBLIOGRAPHY

[1] Gozani, T., "Principles of Nuclear-Based Explosive Detection Systems", in *Proceedings of the First International Symposium on Explosive Detection Technology*, edited by S. Khan ed., New Jersey: Federal Aviation Administration, US Department of Transportation, FAA Technical Center, 27-55, 1992.

[2] Grodzins, L., "Photons In – Photons Out: Non-Destructive Inspection of Containers Using X-Ray and Gamma Ray Techniques", in *Proceedings of the First International Symposium on Explosive Detection Technology*, edited by S. Khan ed., New Jersey: Federal Aviation Administration, US Department of Transportation, FAA Technical Center, 201-231, 1992

[3] Caroli, E., Stephen, J.B., Di Cocco, G., Natalucci, L., and Spizzichino, A., "Coded aperture imaging in x- and gamma-ray astronomy", *Space Science Reviews*, **45**, 349-403, 1987.

[4] Skinner, J.K., "Imaging with Coded-Aperture Masks", *Nuclear Instruments and Methods in Physics Research*, **221**, 33-40, 1984.

[5] Barrett, H.H., and Swindell, W., *Radiological imaging: the theory of image formation, detection, and processing*, New York, Academic Press, 1981.

[6] Fenimore, E.E., and Cannon, T.M., "Coded aperture imaging: many holes make light work", *Optical Engineering*, **19**, 3, 283-289, 1980.

[7] Mertz, L., and Young, N.O., "Fresnel transformation of images", in Proceedings of the International Conference on Optical Instrumentation and Techniques, Chapman and Hall, London, 305, 1961.

[8] Gunson, J., and Polychronopulos, B., "Optimum design of a coded mask x-ray telescope for rocket applications", *Monthly Notices of the Royal Astronomical Society*, **177**, 485-497, 1976.

[9] Barrett, H.H., "Fresnel zone plate imaging in nuclear medicine", *Journal of Nuclear Medicine*, **13**, 6, 382-385, 1972.

[10] Dicke, R.H., "Scatter-hole cameras for X-rays and gamma rays", *The Astrophysical Journal*, **153**, 2, L101-L106, 1968.

[11] Ables, J.G, "Fourier transform photography: a new method for X-ray astronomy", *Proceeedings of the Astronomic Society of Australia*, **1**, 4, 172-173, 1968.

[12] Golay, M.J.E., "Point Arrays Having Compact, Nonredundant Autocorrelations", *Journal of the Optical Society of America*, **61**, 272, 1971.

[13] Fenimore, E.E., and Cannon, T.M., "Coded aperture imaging with uniformly redundant arrays", *Applied Optics*, **17**, 337-347, 1978.

[14] Gottesman, S.R., and Fenimore, E.E., "New family of binary arrays for coded aperture imaging", *Applied Optics*, **28**, 4344-4352, 1989.

[15] Koral, K.F., Freitas, J.E., Leslie Rogers, W., and Keyes, J.W.Jr., "Thyroid scintigraphy with time-coded aperture", *Journal of Nuclear Medicine*, **20**, 4, 345-349, 1979.

[16] Leslie Rogers, W., Koral, K.F., Mayans, R., Leonard, P.F., Thrall, J.H., Brady, T.J., and Keyes, J.W.Jr., "Coded-Aperture imaging of the heart", *Journal of Nuclear Medicine*, **21**, 4, 371-378, 1980.

[17] Leslie Rogers, W., Koral, K.F., and Knoll, G.F., "Digital Tomographic Imaging with Time-Modulated Pseudorandom Coded Aperture and Anger Camera", *Journal of Nuclear Medicine*, **16**, 5, 402-413, 1975.

[18] Liu, Y.H., Rangarajan, A., Gagnon, D., Therrien, M., Sinusas, A.J., Wackers, F.J.Th., and Zubal, I.G., "A novel geometry for SPECT imaging associated with the EM-type blind deconvolution method", *IEEE Transactions on Nuclear Science*, **NS-45**, 4, 2095-2101, 1998.

[19] Fenimore, E.E., Cannon, T.M., Van Hulsteyn, D.B., and Lee, P., "Uniformly redundant array imaging of laser driven compressions: preliminary results", *Applied Optics*, **18**, 7, 945-947, 1979.

[20] Chen, Y.W., Yamanaka, M., Miyanaga, N., Yamanaka, T., Nakai, S., Yamanaka, C., and Tamura, S., "Three-dimensional reconstruction of laser-irradiated targets using URA coded aperture cameras", *Optics Communications*, **71**, 5, 249-255, 1989.

[21] Durrant, P.T., Dallimore, M., Jupp, I.D., and Ramsden, D., "The application of pinhole and coded aperture imaging in the nuclear environment", *Nuclear Instruments and Methods in Physics Research*, **A 422**, 667-671, 1999.

[22] Lanza, R. C., Accorsi, R., and Chen, G., "CAFNA, Coded Aperture Fast Neutron Analysis: application to contraband and explosive detection", in Third International Topical Meeting on Nuclear Applications of Accelerator Technology, ANS Conference Proceedings, LaGrange Park: American Nuclear Society, 147-154, 1999.

[23] Jansson, P.A., ed., *Deconvolution of Images and Spectra*, San Diego, Academic Press, 1997.

[24] Abramovitz, M. and Stegun, I., eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Applied Mathematics Series 55, 1964.

[25] Calabro, C., and Wolf, J.K., "On the Synthesis of Two-Dimensional Arrays with Desirable Correlation Properties", *Information and Control*, **11**, 537-560, 1968.

[26] Nelson, E.D., and Fredman, M.L., "Hadamard Spectroscopy", *Journal of the Optical Society of America*, **60**, 12, 1664-1669, 1970.

[27] Fenimore, E.E., and Weston, G.S., "Fast delta Hadamard transform", *Applied Optics*, **20**, 17, 3058-3067, 1981.

[28] Kopilovich, L.E., "Construction of nonredundant masks over square grids using difference sets", *Optics Communications*, **68**, 1, 7-10, 1988.

[29] Fleming, J.S., and Goddard, B.A., "An evaluation of techniques for stationary coded aperture three-dimensional imaging in nuclear medicine", *Nuclear Instruments and Methods in Physics Research*, **221**, 242-246, 1984.

[30] Baumert, L.D., *Cyclic Difference Sets*, Berlin, Springer-Verlag, 1971.

[31] Proctor, R.J., Skinner, G.K., and Willmore, A.P., "The design of optimum coded mask X-ray telescopes", *Monthly Notices of the Royal Astronomical Society*, **187**, 633-643, 1979.

[32] Fenimore, E.E., "Large symmetric π transformations for Hadamard transforms", *Applied Optics*, **22**, 826-829, 1983.

[33] Cannon, T.M., and Fenimore, E.E., "Tomographical imaging using uniformly redundant arrays", *Applied Optics*, **18**, 7, 1052-1057, 1979.

[34] Giles, A.B., "Self-supporting perfect masks for 2-D infrared and x-ray imaging", *Applied Optics*, **20**, 17, 3068-3071, 1981

[35] in't Zand, J.J.M., Heise, J., and Jeger, R., "The optimum fraction of coded apertures. With an application to the wide field X-ray cameras of SAX", *Astronomy and Astrophysics*, **288**, 665-674, 1994.

[36] Wild, W.J., "Dilute uniformly redundant sequences for use in coded-aperture imaging", *Optics Letters*, **8**, 5, 247-249, 1983.

[37] Gottesman, S.R., and Schneid, S.R., "PNP – A new class of coded aperture arrays", *IEEE Transactions on Nuclear Science* **NS-33**, 745-749, 1986.

[38] Byard, K., "On self-supporting coded aperture arrays", *Nuclear Instruments and Methods in Physics Research*, **A322**, 97-100, 1992.

[39] Byard, K., "Synthesis of binary arrays with perfect correlation properties – coded aperture imaging", *Nuclear Instruments and Methods in Physics Research*, **A336**, 262-268, 1993.

[40] Gourlay, A.R., and Stephen, J.B., "Geometric coded apertures", *Applied Optics*, **22**, 24, 4042-4047, 1983.

[41] Gourlay, A.R., and Young, N.G., "Coded aperture imaging: a class of flexible mask designs", *Applied Optics*, **23**, 22, 4111-4117, 1984.

[42] Gourlay, A.R., Stephen, J.B., and Young, N.G.S., "Geometrically designed coded aperture masks", *Nuclear Instruments and Methods in Physics Research*, **221**, 54-55, 1984.

[43] Fenimore, E.E., "Coded aperture imaging: the modulation transfer function for uniformly redundant arrays", *Applied Optics*, **19**, 14, 2465-2471, 1980.

[44] Fenimore, E.E., and Cannon, T., "Uniformly redundant arrays: digital reconstruction methods", *Applied Optics*, **20**, 10, 1858-1864, 1981.

[45] Skinner, G.K., and Grindlay, J.E., "Coded masks with two spatial scales", *Astronomy and Astrophysics*, **276**, 673-681, 1993.

[46] Hammersley, A., Ponman, T., and Skinner, G., "Reconstruction of images from a coded-aperture box camera", *Nuclear Instruments and Methods in Physics Research*, **A311**, 585-594, 1992.

[47] Ponman, T., Hammersley, A., and Skinner, G., "Error analysis for a noncyclic imaging system", *Nuclear Instruments and Methods in Physics Research*, **A262**, 419-429, 1987.

[48] Willingale R., Sims, M.R., and Turner, M.J.L., "Advanced deconvolution techniques for coded aperture imaging", *Nuclear Instruments and Methods in Physics Research*, **221**, 60-66, 1984.

[49] Byard, K., and Ramsden D., "Coded aperture imaging using imperfect detector systems", *Nuclear Instruments and Methods in Physics Research*, **A342**, 600-608, 1994.

[50] Brown, C., "Multiplex imaging with multiple-pinhole cameras", *Journal of Applied Physics*, **45**, 4, 1806-1811, 1974.

[51] Skinner, G.K., and Ponman, T.J., "On the properties of images from coded-mask telescopes", *Monthly Notices of the Royal Astronomical Society*, **267**, 518-522, 1994.

[52] Fenimore, E.E., "Coded aperture imaging: predicted performance of uniformly redundant arrays", *Applied Optics*, **17**, 22, 3562-3570, 1978.

[53] Busboom, A., Schotten, H.D., and Elders-Boll, H., "Coded aperture imaging with multiple measurements", *Journal of the Optical Society of America*, **A14**, 5, 1058-1065, 1997

[54] Charalambous, P.M., Dean, A.J., Stephen, J.B., and Young, N.G.S., "Aberrations in gamma-ray coded aperture imaging", *Applied Optics*, **23**, 22, 4118-4123, 1984.

[55] Jayanthi U.B., and Braga, J., "Physical implementation of an antimask in URA based coded mask systems", *Nuclear Instruments and Methods in Physics Research*, **A310**, 685-689, 1991.

[56] McConnell, M.L., Forrest, D.J., Chupp, E.L., and Dunphy, P.P., "A coded aperture gamma ray telescope", *IEEE Transactions on Nuclear Science*, **NS-29**, 155-159, 1982.

[57] Dunphy P.P., McConnell, M.L., Owens, A., Chupp, E.L., and Forrest, D.J., "A balloon-borne coded aperture telescope for low-energy gamma-ray astronomy", *Nuclear Instruments and Methods in Physics Research*, **A274**, 362-379, 1989.

[58] Jupp, I.D., Palmer, M.J., Durrant, P.T.. and Ramsden D., "A Comparison of the Performance of Different Gamma-ray Imaging Systems", *IEEE Nuclear Science Symposium*, 1997.

[59] Gotoh, H., and Yagi, H., "Solid Angle Subtended by a Rectangular Slit", *Nuclear Instruments and Methods*, **96**, 485-486, 1971.

[60] Berrim, S., Lansiart, A., and Moretti, J.-L., "Implementing of maximum likelyhood in tomographical coded aperture", *Proceedings of the International Conference on Image Processing*, **2**, 745-748, 1996.

[61] Smith, M.F., Jaszczak, R.J., Wang, H., and Li, J., "Lead and Tungsten Pinhole Insert for I-131 SPECT Tumor Imaging: Experimental Measurements and Photon Transport Simulations", *IEEE Transactions on Nuclear Science*, **44**, 1, 74-82, 1997.

[62] Smith, M.F., Jaszczak, R.J., and Wang, H., "Pinhole Aperture Design for ^{131}I Tumor Imaging", *IEEE Transactions on Nuclear Science*, **44**, 3, 1154-1160, 1997.

[63] Tenney, C.R., Smith, M.F., Greer, K.L., and Jaszczak, R.J., "Uranium Pinhole Collimators for I-131 SPECT Brain Tumor Imaging", Nuclear Science Symposium, 1998. IEEE, **2**, 1312 –1317, 1998.

[64] Koral, K.F., and Leslie Rogers, W., "Application of ART to Time-coded Emission Tomography", *Physics in Medicine and Biology*, **24**, 5, 879-894, 1979.

[65] Shepp, L.A., and Vardi, Y., "Maximum Likelihood Reconstruction for Emission Tomography", *IEEE Transactions on Medical Imaging*, **MI-1**, 2, 113-122, 1982.

[66] Ohyama, N., Honda, T., and Tsujiuchi, J., "Tomogram reconstruction using advanced coded aperture imaging", *Optics Communication*, **36**, 6, 434-438, 1981.

[67] Ohyama, N., Honda, T., Tsujiuchi, J., Matumoto, T., Iinuma, T.A., and Ishimatsu, K., "Advanced coded-aperture imaging system for nuclear medicine", *Applied Optics*, **22**, 22, 3555-3561, 1983.

[68] Paxman, R.G., Smith, W.E., and Barrett, H.H., "Two Algorithms for Use with an Orthogonal-View Coded-Aperture System", *Journal of Nuclear Medicine*, **25**, 6, 700-705, 1984.

[69] Ito, T., and Fujimura, S., "Improvement on depth resolution and reduction of Poisson noise in coded aperture emission CT", *Proceedings of the International Conference on Image Processing*, **2**, 757-760, 1996.

[70] Smith, W.E., Paxman, R.G., and Barrett, H.H., "Image reconstruction from coded data: I. Reconstruction algorithms and experimental results", *Journal of the Optical Society of America*, **A2**, 4, 491-500, 1985.

[71] Paxman, R.G., Barrett, H.H., Smith, W.E., and Milster, T.D., "Image reconstruction from coded data: II. Coded design", *Journal of the Optical Society of America*, **A2**, 4, 501-509, 1985.

[72] Chiu, M.Y., Barrett, H.H., Simpson, R.G., Chou, C., Arendt, J.W., and Gindi, G.R., "Three-dimensional radiographic imaging with a restricted view angle", *Journal of the Optical Society of America*, **69**, 10, 1323-1333, 1979.

[73] Nugent, K. A., "Coded aperture imaging: a Fourier space analysis", *Applied Optics*, **26**, 3, 563-569, 1987.

[74] Gindi, G.R., Paxman, R.G., and Barrett, H.H., "Reconstruction of an object from its coded image and object constraints", *Applied Optics*, **23**, 6, 851-856, 1984.