SCALABLE COMPUTATIONAL OPTICAL IMAGING SYSTEM DESIGNS

by

Ronan Kerviche

A Dissertation Submitted to the Faculty of the

COLLEGE OF OPTICAL SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2017

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Ronan Kerviche, titled "Scalable Computational Optical Imaging System Designs" and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

____ Date: April 26th, 2017

Amit Ashok

____ Date: April 26th, 2017

Matthew A. Kupinski

____ Date: April 26th, 2017

James T. Schwiegerling

Final approval and acceptance of this dissertation is contingent upon the candidates submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____ Date: April 26th, 2017

Dissertation Director: Amit Ashok

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED : Ronan Kerviche

ACKNOWLEDGEMENTS

I first would like to express my extreme gratitude to Professor Amit Ashok for his continuous support and for sharing his valuable experience. I would surely not have been been able to accomplish any of this work without his thoughtful guidance through these years. I would also like to thank Pr. Matthew Kupinski and Pr. Jim Schwiegerling for serving on my dissertation committee. I am grateful to the Technology & Research Initiative Fund for supporting my work through the 2015-2016 academic year. Thank you to all the professors, faculty and staff at the College of Optical Sciences and elsewhere.

Thanks to my fellow lab-mates in the I^{2}SL group for the stimulating discussions and for creating an exciting work environment, in particular to Dr. James Huang, Yuzhang Lin and Jay Voris. Finally, I would like to thank my parents for their ceaseless moral and emotional support during this lifelong journey.

DEDICATION

À mes parents,

Thérèse et Philippe,

Ma grand-mère,

Marie

Et ma famille.

À la mémoire de mes grands-parents,

Denise, Albert et Jean

Contents

List of Figures

List of Tables

Abstract

Computational imaging and sensing leverages the joint-design of optics, detectors and processing to overcome the performance bottlenecks inherent to the traditional imaging paradigm. This novel imaging and sensing design paradigm essentially allows new trade-offs between the optics, detector and processing components of an imaging system and enables broader operational regimes beyond the reach of conventional imaging architectures, which are constrained by well-known Rayleigh, Strehl and Nyquist rules amongst others. In this dissertation, we focus on scalability aspects of these novel computational imaging architectures, their design and implementation, which have far-reaching impacts on the potential and feasibility of realizing task-specific performance gains relative to traditional imager designs. For the extended depth of field (EDoF) computational imager design, which employs a customized phase mask to achieve defocus immunity, we propose a joint-optimization framework to simultaneously optimize the parameters of the optical phase mask and the processing algorithm, with the system design goal of minimizing the noise and artifacts in the final processed image. Using an experimental prototype, we demonstrate that our optimized system design achieves higher fidelity output compared to other static designs from the literature, such as the Cubic and Trefoil phase masks. While traditional imagers rely on an isomorphic mapping between the scene and the optical measurements to form images, they do not exploit the inherent compressibility of natural images and thus are subject to Nyquist sampling. Compressive sensing exploits the inherent redundancy of natural images, basis of image compression algorithms like JPEG/JPEG2000, to make linear projection measurements with far fewer samples than Nyquist for the image forming task.

Here, we present a block wise compressive imaging architecture which is scalable to high space-bandwidth products (*i.e.* large FOV and high resolution applications) and employs a parallelizable and non-iterative piecewise linear reconstruction algorithm capable of operating in real-time. Our compressive imager based on this scalable architecture design is not limited to the imaging task and can also be used for automatic target recognition (ATR) without an intermediate image reconstruction. To maximize the detection and classification performance of this compressive ATR sensor, we have developed a scalable statistical model of natural scenes, which enables the optimization of the compressive sensor projections with the Cauchy-Schwarz mutual information metric. We demonstrate the superior performance of this compressive ATR system using simulation and experiment. Finally, we investigate the fundamental resolution limit of imaging via the canonical incoherent quasi-monochromatic two point-sources separation problem. We extend recent results in the literature demonstrating, with Fisher information and estimator mean square error analysis, that a passive optical mode-sorting architecture with only two measurements can outperform traditional intensity-based imagers employing an ideal focal plane array in the sub-Rayleigh range, thus overcoming the Rayleigh resolution limit.

Chapter 1

Introduction

1.1 Conventional To Computational Imaging

While humans have engaged in visual depiction of their life environment for millennia, beginning roughly with prehistoric hunting scenes around 40,000 BCE, imaging techniques only started to appear with the development of primitive optics. One of the first device built for that purpose was the *camera-obscura*, which projected the view seen through a simple pinhole into a darkened chamber. Many written records of such construct exist in a period ranging from 500 BCE as it was built for studying light, to the 11th century for its use as scientific instrument and entertainment. Of course, optics would continue to improve over the next centuries and take giant leaps at times, with for instance Lord Newton's design of the first reflective telescope in 1668. Yet, while these image-forming devices were developing, the underlying reproduction process was primarily manual. The birth of modern photography happened at the beginning of the nineteenth century with the pioneering contributions of Nicéphore Niépce and Louis Daguerre. They used of a photosensitive surface to record light via a slow chemical process requiring a full day of exposure to imprint. Nearly a century later, the invention of the camera tube, the first photoelectric image sensor [1, 2, 3], advanced the recording process to the next generation. Combined with Cathode Ray Tube displays, that system enabled the capture and rendering of video signals which could be transmitted through radio-waves and gave rise to first television broadcastings in the 1930s.

The most recent image sensor revolution started in 1969 with the invention of the Coupled Charge Device at AT&T Bell Labs by Willard Boyle and George E. Smith [4] which enabled the digitalization of images and their processing with computers. The democratization of electronic cameras was further accelerated with the development of Active Pixel Sensor in the 1990s, with better energy efficiency and lower price points due to the integration with silicon processes. With these advances, imagers are presently not only employed for visual communication but also in numerous applications via machine vision, such as autonomous driving, surveillance and security, medical diagnostics and natural user interfaces. This trend has notably benefited from the exponential growth of computing capabilities, following Moore's law [5], and enabled the progress of complex algorithms for image processing and analysis.

However, it is interesting to note that the overall architecture of optical imagers in the visible and infrared domains has remained essentially unchanged during the last century. While more complex and precise optical objectives have been developed over the years, for instance zoom lenses or apo- and super-apochromats, the design and implementation of these image-forming optics subsystems has been largely isolated from the image sensor and post-processing stages. Thus, the primary imaging performance criteria still remains an absence of prominent aberrations and distortions and tend towards diffraction-limited performance. Nevertheless, this isomorphic approach to imaging, and more broadly sensing, is facing greater challenges in recent years as the general performance is becoming increasingly limited by the physical and design constraints rather than technology. This is amply demonstrated for instance by the integration of high resolution cameras into sub centimeter-scaled packages to be integrated in smart-phones and other handheld devices, where the field curvature and the low F-Number effectively limit the final image quality [27]; or for example in hyper-spectral imaging where the signal readout bandwidth imposes a severe limitation on the scalability of the system toward higher space-spectral-bandwidth products.

In contrast to such isomorphic imaging system architectures, other imaging systems require an algorithm to reconstruct the desired image from the measurement data because of the indirect (non-isomorphic form) mapping from object to data measurement space. Such indirect imaging systems are commonly employed for non-invasive medical imaging techniques based on tomography, such as X-Ray, MRI, PET. In this indirect imaging domain, an extensive corpus of research presents a multitude of methods to assess of the resulting image quality and its impact on future exploitation [6, 7, 8]. This approach provides a relevant example of an end-to-end system performance analysis.

Computational imaging (CI) breaks away from conventional imaging paradigm which considers optics, detection and processing elements of the system in isolation. The goal of the CI paradigm is to enable analysis and design of the three elements jointly in order to potentially achieve a global optimal design and performance, not always achievable by the traditional imaging approach. The pioneering work on an extended depth of field (EDoF) imager is an ideal example of the computational imaging design. This system design employs a mask to induce depth-independent aberrations in the optical domain. However, despite the reduction in the magnitude of the modulation transfer function, the system becomes nearly invariant to the scene depth. Therefore, it is possible using a post-measurement digital filter to boost the modulation function and deliver a nearly ideal in-focus image over a broader range of object conjugates relative to a traditional imager with the same F-number and focal length. In order to produce the same depth-of-field extension within the traditional imaging paradigm one would need to stop down (or apodize) the aperture, which degrades both the photon throughput and the diffraction limited resolution. Hence the elegant, and unconventional EDoF imager design can achieve an operating point beyond the reach traditional imagers.

The EDoF computational imager design illustrates how balancing between optics and post-processing design degrees of freedom enables improved system performances. Such a joint-design approach can be used to achieve other system performance and complexity trade-offs not accessible by traditional isomorphic imaging. Optical compressive sensing (CS) is another noteworthy computational imaging architecture which relies on the inherent sparsity/compressibility of natural scenes to reduce the measurement sampling requirement below the Nyquist rule. The underlying redundancy and hence the compressibility of natural images is evident by the widespread success of image compression algorithms, such as JPEG/JPEG2000. However, conventional imaging approach employs isomorphic measurements and thus ignores the inherent redundancy of natural scenes, leading to the measurement of redundant data. As a result, conventional imagers incur several penalties such as performing many times more measurements than required, having larger power consumption and higher complexity FPA. In compressive imaging, the measurements correspond to linear projections of the optical image onto some compact basis, such as the discrete cosine transform (DCT), discrete wavelet transform (DWT) or random projections, and requires relatively fewer measurement in proportion to the underlying sparsity of natural images. Such optical pre-processing (projection) before detection has several inherent advantages such as : fewer detectors and therefore lower complexity FPA, lower readout bandwidth, lower power consumption, lower heat and multiplexing gain advantage for thermal-noise limited detectors. As the compressive imaging measurements are not isomorphic samples but instead projections of the optical image they need to be processed by a reconstruction algorithm to form an image.

Another key aspect of computational imaging is its inherent suitability to support specific tasks, such as target classification, target tracking, *e.t.c.* The inherent joint-design nature of computational imaging allows for a logical incorporation of specific tasks in the post-processing algorithm which in turn determine the form and function of front-end optics, so as to maximize the system-end-to-end performance using an objective task-specific metric. Machine vision related tasks, such as target recognition is a pertinent example of task specific imaging, where an intermediate image of the scene is not required, but only the scene-related data for direct target recognition. For such tests, the appropriate system performance is the probability of error, while the mean square error metric is commonly employed for traditional image formation tasks.

1.2 Main Contributions

While computational imaging paradigm alleviates some of the burdens of imager design and performance bottlenecks inherent in traditional imaging architectures, it encompasses its own set of design and implementation challenges. Given the different design approaches of computational imaging and traditional imaging, the former typically involves higher system complexity to achieve performance improvement relative to lower system complexity and lower threshold of integration between optics, measurements and processing for the latter. In this body of research we examine the scalability of computational imaging sensor designs. For traditional imagers, simple design properties determine the scalability to higher space-bandwidth (*i.e.* to increase the angular resolution and/or the field of view) or temporal-bandwidth (*i.e.* to increase the frame-rate). In computational imaging and sensing, this notion of system scalability extends beyond their system performance parameters, and include other complex design aspects such as processing, scene models and system optimization metrics. More specifically, given the integral nature of post-measurement algorithm, the computational complexity of algorithm implementations becomes a critical component in system scalability. Similarly, the underlying scene models that are required for system optimization and processing need to be scalable with increasing complexity of the exploitation task and to high space-bandwidth products. Thus it becomes imperative that such challenges be addressed, in part by high-performance computational architectures such as distributed processors and graphics processing units (GPUs). To address these designs and implementation challenges, we construct efficient scene models and tractable system design optimization frameworks, to design, build and test imagers and prototypes.

The main contributions of this dissertation are as follows :

- A joint-optimization framework is described for the design of an EDoF imager. The system design metric provides a measure of the end-to-end system performance rather than separately considering the performance of the optics, sensing and processing stages separately. In our work, we also employ higher order Zernike polynomials demonstrating scalability in design complexity that contributes to a significant performance improvement. The resulting phase mask design coupled with image restoration algorithm produces higher fidelity image output due to lower noise amplification relative to other phase masks design (cubic, trefoil) as well as providing a longer depth range than a traditional imager. By leveraging a GPU-based implementation of physics-based system forward modeling and post-measurement restoration algorithm we were able to explore the full decision space represented by the large optical and algorithm design degrees of freedom. To validate our joint-design approach, we optimized and constructed a germanium phase mask-based F/1 EDoF long-wave imager with an infrared camera and quantified its performance via extensive simulations and lab experiments.

- Design and implementation of a programmable compressive imager testbed that is inherently scalable to large space-bandwidth product. The compressive imager acquires high-resolution images using a low-resolution focal plane array (FPA) while collecting 4x to 8x fewer measurements relative to a conventional imager sampling at Nyquist rate. We have developed a fully automated method for system calibration to minimize deviations between the ideal system design and its imperfect optical implementation. We have also developed a non-iterative piecewise linear estimation algorithm for image reconstruction from compressed measurements that is scalable to higher frame-rates by exploiting parallel implementation. With this testbed we were able to experimentally verify the superior performance of information optimal projections in comparison to several common random projection patterns relative to the random projection patterns widely used in the compressed sensing community.

- In addition to the image formation task, we also demonstrate how this programmable compressive imager testbed can be used to implement automatic target recognition (ATR), without having to reconstruct an intermediate image. We consider the task of detecting and classifying specific targets against a cluttered background. We have developed a scalable scene model that accounts for typical variations in the scene, including background and target appearance (rotation, location, scale). Using the scalable scene model and a computationally tractable information theoretic measure (CSMI) we optimize the linear projection basis to maximize the relevant target information. In contrast to well-known information measures, such as Shannon's Mutual Information, we can express both the objective function, *i.e.* CSMI, and its gradient in closed form for our Gaussian mixture likelihood scene model. Furthermore, we show that CSMI provides an upper-bound on the probability of misdetection and misclassification for the system. Using this scalable information theoretic system design framework, we demonstrate with simulation and experiment that the information optimized compressive projections outperform random projections and other compressive designs as well as the conventional imager in low SNR (0 to 10dB) and high-compression (between 42x and 64x) regimes.

- We analyze the canonical incoherent and quasi-monochromatic point-sources separation problem from an information theoretic perspective, given its relevance in defining the incoherent resolution of an optical imager system. Prior work by Tsang *et al.* has already demonstrated that a mode decomposition of the optical field in the image plane capture more of the Fisher Information than a traditional imager employing an ideal FPA. This is rather a surprising result given that per Rayleigh's criterion, the estimation error of angular separation degrades quickly below the angular diffraction limit, while the mode-sorting measurement can achieve constant error performance, including in the sub-Rayleigh regime. Our contribution in the analysis and design of mode-sorting measurement for the two point-sources resolution problem includes extending the mode-design to hard aperture, as well as generalizing a two-modes measurement design to any arbitrary aperture, and showing that any known optical aberrations in the imager pupil can be perfectly compensated by the appropriate mode-basis design. Finally, we analyze two candidate measurement architectures in the sub-Rayleigh range and their angular estimation performance in presence of implementation non-idealities.

1.3 Dissertation Organization

The rest of this dissertation is organized as follows. In chapter 2 we describe a joint-optimization framework for the design of an EDoF imager in thermal infrared spectral regime for a set of system specifications, including the F-Number, depth range and maximum acceptable noise gain. We employ a system transfer function (STF) to capture the system's end-to-end performance. The optical phase mask is parametrized in terms of the Zernike polynomials, whose coefficients serve as system design parameters and are optimized subject to system specifications. The final system output for the optimized EDoF phase mask design is compared to other existing designs in the literature such as the cubic and trefoil profiles. The EDoF system prototype is constructed and its performance is tested against a conventional imager. A scalable block wise compressive imager testbed is discussed in chapter 3, which has notable advantages over both conventional and single pixel camera architectures in terms of scalability with respect to space-bandwidth products. The random and information optimal compressive projection designs are tested on this programmable testbed to assess their respective output image quality. Chapter 4 describes how this compressive imaging testbed can be used to implement an Automatic Target Recognition task, without intermediate image reconstruction, by designing appropriate compressed measurements. We employ the Cauchy-Schwarz mutual information metric coupled with a scalable scene model to design compressive projection patterns for this specific task. Using simulation and experimental data, we demonstrate the effectiveness of the proposed task-specific imager design. In chapter 5, we pursue a theoretical analysis of the angular resolution limit to an optical system with a finite aperture using the canonical two-point resolution problem. We quantify the performance of an imager with a focal plane array measurement using Fisher information metric and verify Rayleigh's resolution limit. We then extend the analysis of an alternate mode-sorting measurement design by Tsang *et al.* to a hard aperture and generalize a two-mode design analysis to any arbitrary aperture. We conclude in chapter 6 with some key observations of this body of research and directions for future work.

Chapter 2

Joint-Design Approach for Extended Depth of Field Imaging

While a shallow depth of field is often used to emphasize the subject of a photograph, it can also be a burden in applications constrained to use fast optics to compensate for low scene flux and/or low SNR. The computational extended depth of field (EDoF) imager solves this problem by introducing specific aberrations in the optical train which neutralize the defocus and can be corrected by digital processing. In this chapter, we propose a joint-design and optimization framework which optimizes both the parameters of the optical phase mask and of the digital restoration algorithm to maximize the overall image quality delivered by the computational imager.

2.1 Introduction

The depth of field of an optical imaging system is an important operational parameter in many applications such as microscopy and photography. Traditional optical approaches to control the depth of field rely on changing the f-number (f/#) or numerical aperture (NA) and/or aperture apodization, which can potentially degrade the light collection and the resolution of an imager significantly. Such system performance trade-offs inherent in traditional imaging system designs can be overcome by considering optical and computational design degrees of freedom jointly within a computational imaging framework. A point in case is the pioneering computational optical imaging, proposed by Dowski and Cathey [9], which extends the depth of field (DoF) of an imager by employing a cubic phase mask in conjunction with linear post-measurement processing. This non-traditional imaging approach is especially suitable for applications such as microscopy and surveillance where an extended depth of field (EDoF) is desirable without concomitant signal to noise ratio (SNR) penalty incurred with traditional optical-only approaches. Since this pioneering EDoF imager design with the cubic phase mask, numerous improvements with alternate phase masks designs have been proposed in the literature. Some of the notable EDoF phase mask designs include, the canonical polynomial basis based phase mask [21], the Log-Asphere phase mask [24], the exponential phase mask [26], the fractional power and binary amplitude phase mask [25], the freeform mesh mask [23]. It is noteworthy that nearly all of these phase mask designs rely on purely optical metrics, such as Strehl ratio, optical point spread function (PSF) or modulation transfer function (MTF) defocus invariance, ambiguity function, and therefore, not truly joint designs because they ignore the end-to-end system performance that includes the computational subsystem.

There have been a few EDoF phase mask designs that have indeed considered the joint end-to-end system optimization, including the work of [20, 22]. However, such designs only include relatively few optical design degrees of freedom and as such are unable to leverage the full potential of the computational imaging paradigm. To our knowledge there has been no comprehensive phase mask design study for EDoF imaging over a large set of optical design degrees of freedom examining the inherent trade-off between measurement signal-to-noise ratio (SNR) and DoF extension. While in Refs. [18], [19], S. Bagheri *et al.* present a theoretical framework that yields a bounding expression for the inherent trade-off between the extension DoF and measurement SNR budget, they do not provide actual phase mask designs that achieve or even approach these theoretical limits. In this work, we propose a framework to jointly optimize a circular phase mask using an end-to-end system modulation transfer function (STF) based metric to achieve a desired DoF extension, given a finite SNR budget. We consider the powerful Zernike parametrization of the optical phase mask over a large set of polynomial coefficients that allows access to a rich design space. Our choice of Zernike polynomial representation is based on an efficient mathematical representation of an optical surface on a circular aperture. Furthermore, the Zernike polynomials are routinely used in optical design due to its relation to various aberrations (e.g. Seidel aberrations). We consider the application of this joint design framework to an EDoF imaging system design and experimental validation in the long wave infrared (LWIR) spectral band.

2.1.1 Limitation of Traditional Approach

The DoF of a traditional imaging system can be maximized by focusing it at the hyperfocal distance s_{HF}. At this focal distance, all object conjugates between half of the hyperfocal distance up to infinity are *in focus*, in the paraxial regime. The hyperfocal distance s_{HF} can be expressed as a function of imaging system's focal length f, its F-Number $F_\sharp$ and the admissible blur size c : $s_{HF} \approx f^2/(cF_\sharp)$. Therefore, the DoF at the hyperfocal distance can be extended by simply decreasing the f/# (i.e. slower f/#) of the imaging system. Note that such a DoF extension approach is also applicable for other focus positions as well. However, as mentioned earlier this traditional approach incurs two performance penalties : (a) reduced light collection leading to lower measurement SNR, and (2) degradation of diffracted-limited resolution. The relative SNR loss can be quantified per Equation 2.1, where $f/\#_{ref} < f/\#_{slow}$, as illustrated in Figure 2.1. The impact of slower f/# on the diffraction-limited resolution (at $\lambda = 10\mu m$) is also evident in Figures Figure 2.2. This motivates a non-traditional approach to extending depth of field by jointly exploiting the optical and the computational degrees of freedom, as outlined in the introduction section.

$$SNR_{loss} = 10\log_{10}\left(\frac{f/\#_{slow}}{f/\#_{ref}}\right)^4 \tag{2.1}$$

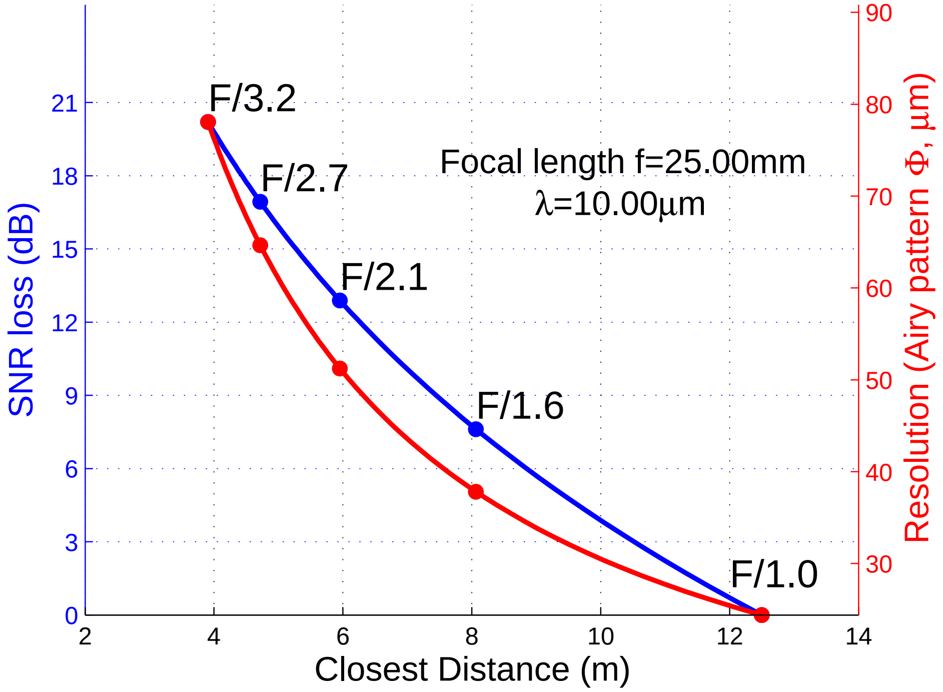


Figure 2.1: Standard imaging system characteristics : SNR loss and PSF diameter as functions of the F-Number/closest in-focus distance.

2.2 Joint System Design Framework

In this section, we describe our joint imaging system design framework that encompasses a high-order Zernike phase mask, a linear image restoration algorithm, an end-to-end system design metric and the system optimization method. We begin by defining our optical system model.

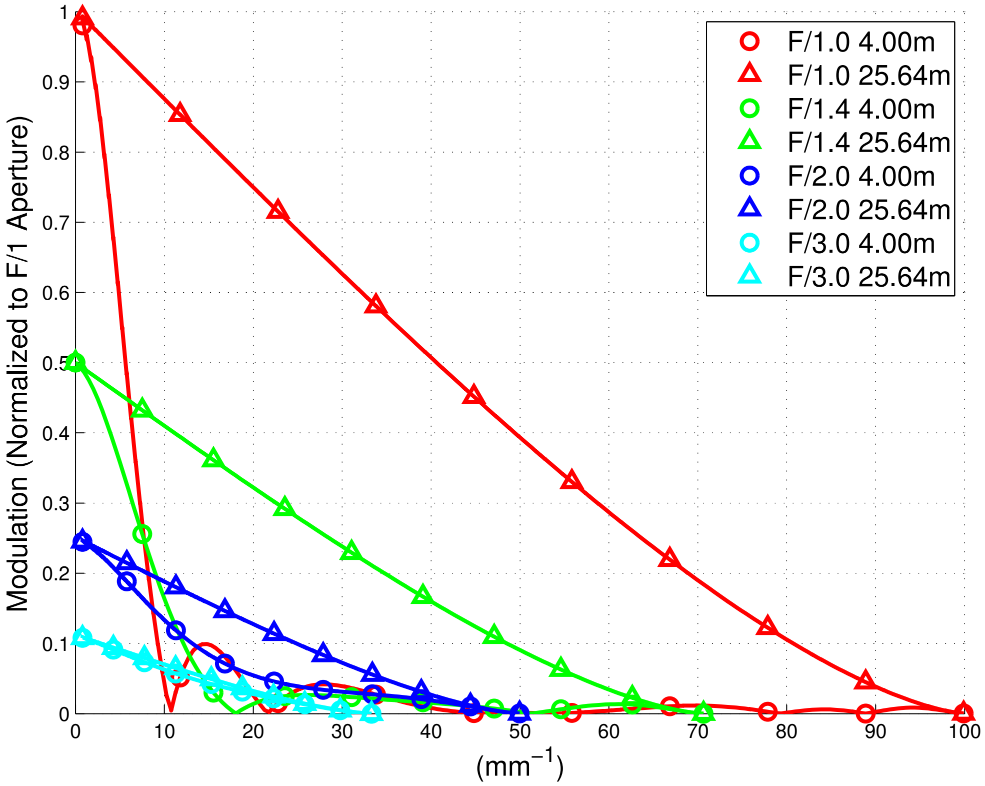


Figure 2.2: Comparison of the diffraction limited and defocused MTFs induced by a point source at the 25m (the hyperfocal distance in this context) and 4m respectively. The modulation functions are normalized to the largest aperture ($F/1$) to illustrate the energy lost while stopping down the aperture.

2.2.1 Optical System Model

The optical component of the computational imaging system is modeled by its thin-lens equivalent, which yields an accurate optical PSF description in the paraxial regime. Given this model, we can express the wavefront in the exit-pupil $U(x, y, s_o)$ in terms of the aperture function $P(x, y)$, the phase mask function $\boldsymbol{\alpha}$, and object/image distance s_o/s_i :

$$U(x, y; s_o) = U_0 \; P(x, y) \; \exp\left(j\frac{\pi}{\lambda} \left[\left(\frac{1}{s_o} + \frac{1}{s_i} - \frac{1}{f} \right) (x^2 + y^2) - 2\Delta_{\boldsymbol{\alpha}}(x, y) \right] \right). \tag{2.2}$$

The optical path length (OPL) function $\Delta_{\boldsymbol{\alpha}}(x, y)$ associated with the the phase mask profile $S_{\boldsymbol{\alpha}}(x, y)$ is defined as: $\Delta_{\boldsymbol{\alpha}}(x, y) = (n_o - 1)S_{\boldsymbol{\alpha}}(x, y)$, where n_o is the refractive index of the phase mask material. This wavefront expression, in the exit-pupil, leads to the following definition of the (complex) optical transfer function (OTF) in Equation 2.3, where $\star$ denotes the autocorrelation operator.

$$OTF(\xi, \eta; s_o) = \frac{[U \star U](\lambda \xi s_i, \lambda \eta s_i; s_o)}{[U \star U](0,0)}. \tag{2.3}$$

The defocus phase $\Psi(R)$ is defined as :

$$\Psi = \frac{R^2}{2\lambda}\left(\frac{1}{f} - \frac{1}{s_o} - \frac{1}{s_i}\right), \tag{2.4}$$

where $R = \sqrt{x^2 + y^2}$ is the radial pupil coordinate.

2.2.2 Zernike Phase Mask Parameterization

We parameterize the phase mask surface in terms of Zernike polynomials, which are orthonormal over the unit circle and widely used in metrology to describe system aberrations that arise due to departure from the reference sphere. Thus the phase mask profile $S_{\boldsymbol{\alpha}}(\rho, \varphi)$ can be defined in terms of the first N Zernike polynomials $Z_{k:\{m;n\}}(\rho, \varphi)$ in Equation 2.5, where α_k denotes the expansion coefficient for the k^{th} polynomial and ρ/φ are the polar coordinate variables. The set of parameters $\{\alpha_1, \ldots \alpha_N\}$ or parameter vector $\boldsymbol{\alpha}$ represents the optical design degrees of freedom.

$$\Delta_{\boldsymbol{\alpha}}(\rho, \varphi) = \sum_{k}^{N} \alpha_k Z_{k:\{m;n\}}(\rho, \varphi). \tag{2.5}$$

The n^{th} order Zernike polynomials $Z_n^m(\rho, \varphi)$ are defined as follows :

$$\forall n \geq 2,\ |m| < n,\ (n-m) \in \{2p; p \in \mathbb{Z}\},$$

$$Z_n^m(\rho,\varphi) = \begin{cases} R_n^m(\rho)\cos(m\,\varphi), & \text{if } m > 0 \\ R_n^m(\rho)\sin(m\,\varphi), & \text{otherwise} \end{cases} \tag{2.6}$$

$$R_m^n(\rho) = \sum_{k=0}^{(n-m)/2} (-1)^k \binom{n-k}{K} \binom{n-2k}{(n-m)/2-k} \rho^{n-2k}, \quad \rho \in [0;1]. \tag{2.7}$$

The second and third order Zernike polynomials are illustrated in Figure 2.3.

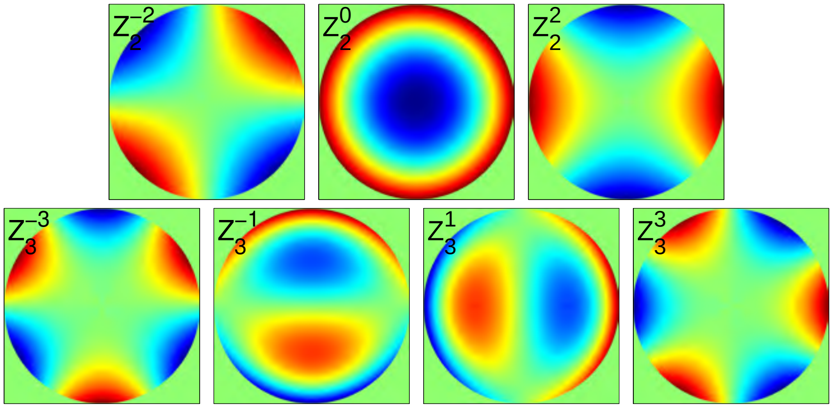


Figure 2.3: Zernike Polynomials of order $n = 2$ (*top-row*) and $n = 3$ (*bottom-row*). From left to right, first row : *astigmatism*, *defocus*, *astigmatism*; second row : *trefoil*, *coma*, *coma*, *trefoil*.

It is instructive to note that any radial profile through the center (i.e. optical axis) of an even order Zernike polynomial is of course an even function. Thus the corresponding wavefront is locally similar to a defocus wavefront. This indeed also holds for any linear combination of even order polynomials with arbitrary weights, due to the properties of even functions. Thus, we can infer that all even order polynomial terms primarily introduce some order of local focusing of the wavefront. This points to the inherent design redundancy associated with the parameter α_k of all even order polynomials parameters, as such we set all such parameter values to zero, except for the main defocus term $Z_2^0(\rho) = 2\rho^2 - 1$. The rationale for including the main defocus it that it serves to center the EDoF response within a desired range of object conjugates.

Overall, the Zernike parameterization of the phase mask is key to the performance improvements that derive from the EDoF imaging system design described in this work. We believe that this is primarily due to two reasons : (1) Zernike polynomials offer an efficient representation of functions (i.e. wavefronts) on the unit circle that corresponds to a circular aperture that is relevant in most optical imaging systems, and (2) such an efficient representation enables inclusion of a larger number of optical parameters that extends the system design space. Furthermore, the direct correspondence between the Zernike polynomials and the Seidel aberrations also provides guidance towards the inclusion or exclusion of particular order(s) of the Zernike polynomial, as discussed in the preceding paragraph. It is important that point out that while any description of a phase mask in Zernike polynomials can be equivalently expressed in any alternative complete bases (e.g. polynomials in the Cartesian coordinate), however, such an alternative description likely leads to a larger number of parameters. This would lead to an inefficient phase mask representation and as a result thereof increased complexity of phase mask design/optimization. Thus, the efficiency of the phase mask representation, with respect to the computational complexity of system design, becomes an important consideration towards the ability to explore a larger design space.

2.2.3 Image Reconstruction Algorithm

To incorporate the optical PSF into the post-measurement image reconstruction stage of the computational imaging system, we choose the Wiener filter Equation 2.8 that is the optimal linear filter design in the mean square error (MSE) sense [15]. Moreover, the Wiener filter enables the inclusion of statistical model of scene variability via its power spectral density (PSD) and accounts for the finite image sensor noise in terms of the signal to noise ratio (SNR) W_{SNR}, thereby potentially improving the reconstruction performance. Given a target or a reference object distance s_w at which we want to achieve perfect restoration of the optical PSF or OTF, up to the diffraction-limited performance, the Wiener filter $W_{\boldsymbol{\alpha},s_w}(\xi,\eta)$

can be expressed as :

$$W_{\boldsymbol{\alpha},s_w}(\xi,\eta) = \frac{OTF_{\boldsymbol{\alpha}}(\xi,\eta,s_w)^* \; \Gamma_{\text{envelope}}(\xi,\eta)}{|OTF_{\boldsymbol{\alpha}}(\xi,\eta,s_w)|^2 + S_{\text{noise}}(\xi,\eta)/S_{\text{model}}(\xi,\eta)}, \tag{2.8}$$

where $\Gamma_{\text{envelope}}(\xi,\eta)$ is the target diffraction-limited OTF, $S_{\text{noise}}(\xi,\eta)$ is the PSD of the image sensor noise, and $S_{\text{model}}(\xi,\eta)$ is the scene PSD model. The scene and image sensor noise PSD models are defined as follows :

$$S_{\text{model}}(\xi,\eta) = \frac{\gamma}{(\xi^2+\eta^2)^{\beta/2}} \qquad \text{and} \qquad S_{\text{noise}}(\xi,\eta) = 10^{-W_{\text{SNR}}/10}, \tag{2.9}$$

where the scalar parameters β and γ are obtained by fitting the average PSD of a dataset of in-focus grayscale images.

In applications where the measurement noise is significant, such as micro-bolometer based imaging sensor in LWIR, the noise gain G_n is an important system performance metric. It is defined as the integral of the amplification power over the OTF domain $\mathbf{P}$:

$$G_n = 10\log_{10}\left(\iint_{\mathbf{P}} |W_{\boldsymbol{\alpha},s_w}(\xi,\eta)|^2 \, d\xi d\eta \Big/ \iint_{\mathbf{P}} d\xi d\eta\right). \tag{2.10}$$

In the Wiener filter design, the parameter W_{SNR}, defined as the measurement SNR, determines the trade-off between the OTF restoration and the noise gain.

2.2.4 System Design Metric and Optimization

The system design metric $g(\boldsymbol{\alpha})$, which is a measure of the system transfer function (STF) departure from a reference flat (unity) profile, is defined as follows,

$$g(\boldsymbol{\alpha}) = \int_{s_i^{\min}}^{s_i^{\max}} \iint_{\mathbf{P}} |1 - STF_{\boldsymbol{\alpha}}(\xi,\eta,s_o(s_i))| \; ds_i d\xi d\eta, \tag{2.11}$$

where the STF is defined as the product of the OTF and Wiener image reconstruction filter profile :

$$STF_{\boldsymbol{\alpha}}(\xi, \eta, s_o) = OTF_{\boldsymbol{\alpha}}(\xi, \eta, s_o) \cdot W_{\boldsymbol{\alpha}, s_w}(\xi, \eta). \tag{2.12}$$

Thus the overall goal of this system design metric is to minimize the STF variations over a desired depth of field across all spatial frequencies. In other words, achieve defocus invariance across a depth of field. There are some subtle but important considerations that motivate the choice of this particular system design metric. First, a design metric based on the STF captures the end-to-end system performance, which is consistent with a joint system design approach. Second, it is also important to note that while the ideal STF profile should take the form of a diffraction-limited OTF, the reference flat profile is chosen *instead* to emphasize the restoration of the high spatial frequencies, thus maximizing the overall system resolution. Furthermore, the choice of the $\mathcal{L}_1$ distance in the system design metric formulation, as opposed to say a $\mathcal{L}_2$ distance, is to measure departure from flat reference provides a higher sensitivity to OTF amplitude variations across the depth of field. For example, OTF amplitude variations around a low modulation value can be more deleterious to the system performance compared to variations around higher modulation values. This is due to the action of Wiener image reconstruction filter whose gain is inversely proportional to the OTF amplitude (modulation). Therefore, in effect this system metric behavior promotes phase mask designs that yield relatively constant OTF amplitude across the depth of field with higher amplitude, thus simultaneously minimizing the noise gain and reducing reconstruction artifacts that derive from OTF variations through focus. Finally, we note that the computation of the system design metric in Equation 2.11, employs numerical integration over spatial frequencies ξ/η. More importantly, the remaining outer integral in Equation 2.11 defined over a range of image distances $[s_i^{\min}, s_i^{\max}]$ instead of equivalent range of object distances $[s_o^{\min}, s_o^{\max}]$. This choice of integration over image space, compared to the object space, results in a higher numerical accuracy for a fixed number of integrand evaluations because it samples the closer (to lens) object conjugates more densely where the defocus and hence the integrand changes rapidly.

Given the non-convex nature of the system design metric coupled with the relatively high dimensionality of the search space (equal to the number of coefficients chosen for the phase mask description), we employ simulated annealing [16], a global optimization technique, for system optimization. As simulated annealing does not guarantee global optimal for a finite number iterations, we repeat the optimization process, each with a random starting point, a number of times to obtain a set of solutions. Each optimization run involves exploration of approximately 100,000 candidate systems. As we discuss in depth in the next section, we find that there is no unique system design with the best performance, but rather a set or family of system designs with visually similar phase masks.

Focal length	25.0 mm
$F_\sharp$	1.0
Center wavelength	10.0 μm
Pixel pitch	25.0μm
Optical cut-off frequency	$\approx 100 mm^{-1}$
Sensor cut-off frequency	$20mm^{-1}$
Hyperfocal distance in object space	25.6 m
HF/2 or closest *acceptable* distance	12.8 m
Target Closest distance	4.0 m
SNR_{loss} after increassing the F-Number of a standard system	19.7 dB
Maximum tolerable noise gain (SNR_{loss})	3.0 dB

Table 2.1: Characteristics of the targeted optical system

2.3 System Design Results and Analysis

In our design study we pursue optimization of three different EDoF imaging systems within the proposed joint design framework : (1) a traditional cubic phase mask with a third-order phase term and a defocus term: $\Delta_{\text{cubic}}(x,y) = \alpha_1(x^3+y^3)+\alpha_2(x^2+y^2)$, (2) a trefoil phase mask design, with two trefoil Zernike polynomial terms along with a defocus term: $\Delta_{\text{trefoil}}(\rho,\varphi) = \alpha_{(3,-3)}(\rho^3\sin(3\varphi)) + \alpha_{(3,+3)}(\rho^3\cos(3\varphi)) + \alpha_{(2,0)}(2\rho^2-1)$, and (3) our proposed Zernike-based phase mask, with only the odd orders (up to $n=7$) and the main ($n=2$ order) defocus terms : $\Delta_{\text{Zernike}}(\rho,\varphi) = \sum\sum_{m,n=odd}\alpha_{m,n}Z_n^m(\rho,\varphi)+\alpha_{2,0}(2\rho^2-1)$. The system design parameters and specification for this study are listed in Table 2.1. We will use the long wave infrared (LWIR) spectral band ($\lambda = 10\mu m$) to illustrate the joint design approach in this work. The desired depth of field for this system design extends from $4m$ out to $+\infty$ that corresponds to a defocus range of $\Phi = 1.65\lambda$. The maximum noise gain budget is set to $3dB$ to limit the noise amplification in the image reconstruction. Note that these design targets are meant to be illustrative, and in principle the proposed joint design framework and the Zernike phase mask design can be applied to any optical spectral band and depth of field requirements.

The optimized optical phase mask prescriptions for each of the three designs are tabulated in Table 2.2, Table 2.3, and Table 2.4 and their surface interferograms are shown in Figure 2.16. It is interesting to note that the optimized defocus term is similar across all the three phase mask designs. This illustrates that the optimization process effectively adjusts the original focus point (hyperfocal distance) to a closer object distance to achieve a symmetric defocus range across the desired depth of focus. The corresponding two-dimensional (2D) MTFs of the three optimized EDoF systems are shown in Figures 2.5 and 2.6 for two different object distances. Figure 2.4 show the one-dimensional profiles of the 2D MTFs in X (horizontal) and XY (diagonal) directions. Note that even though the cubic phase mask phase profile is separable in x and y dimensions, the circular aperture renders it non-separable, as evident in these one-dimensional MTF profiles.

For the conventional imaging system we observe a significant loss of modulation for object ranges between $4m$ and $10m$ as shown in Figure 2.4. In comparison, the cubic phase mask design achieves a significant reduction in modulation variance across focus. However, the modulation across the diagonal direction (XY) is significantly degraded relative to horizontal direction (X). This leads to a STF that has excess modulation (i.e. greater than unity) corresponding to the spatial frequencies where the OTF amplitude is relatively small. Such large variations in the STF across the focus leads to significant artifacts (see images in Figures 2.10 to 2.15). The optimized trefoil phase mask design improves upon the cubic phase mask design by implementing a more symmetric phase profile that includes both elements (i.e. positive and negative orders) and achieves nearly identical OTF along X and XY directions. Furthermore, the resulting STF is relatively well behaved compared to the cubic phase mask. The optimized Zernike phase mask design achieves a higher degree of OTF invariance, given the additional DoF, which yields a well behaved STF across the desired DoF.

Note that the reconstructed modulation transfer profiles are embedding prior knowledges of the average power spectrum of natural scenes and noise from the Wiener filter expression. This tends to affect the high spatial frequency range.

We also tested free optimization of the all the Zernike coefficients up to the 7th order and the optimization process tends to validate our assumption over the contribution of the even orders of Zernike polynomials ($n = 2$). After the optimization process, the coefficients of the corresponding polynomials are low, in magnitude, compared to other significant weights. They are usually non-zero because of the random fluctuations introduced by the Simulated Annealing algorithm. The only coefficient not following this observation is the defocus term which is always kept at the previously discussed value. However, the stochastic optimization process does not return here a single solution to the problem but rather solutions belonging to two different sets depending on if the phase mask is primarily based on coma or trefoil Zernike components, as shown in Figure 2.17.

ID	**Cartesian Coefficient**	**(in unit of λ)**	**Expression**
1	Defocus (n=2)	0.629	$(x^2 + y^2)$
2	Cubic (n=3)	1.934	$(x^3 + y^3)$

Table 2.2: Cubic phase mask coefficients (for the normalized radial coordinate : $\rho = \sqrt{x^2 + y^2} \leq 1.0$).

ID	**Zernike Coefficient**	**(in unit of λ)**	**Expression**
1	Defocus (n=2, m=0)	0.313	$(2\rho^2 - 1)$
6	Trefoil 1 (n=3, m=-3)	0.700	$(\rho^3)\sin(3\varphi)$
7	Trefoil 2 (n=3, m=3)	0.760	$(\rho^3)\cos(3\varphi)$

Table 2.3: Trefoil phase mask coefficients.

The reconstruction step is adjusted so that the noise-gain generated by the linear Wiener filter is equal to $3.0dB$ for the novel design; this is managed by adjusting the single Wiener

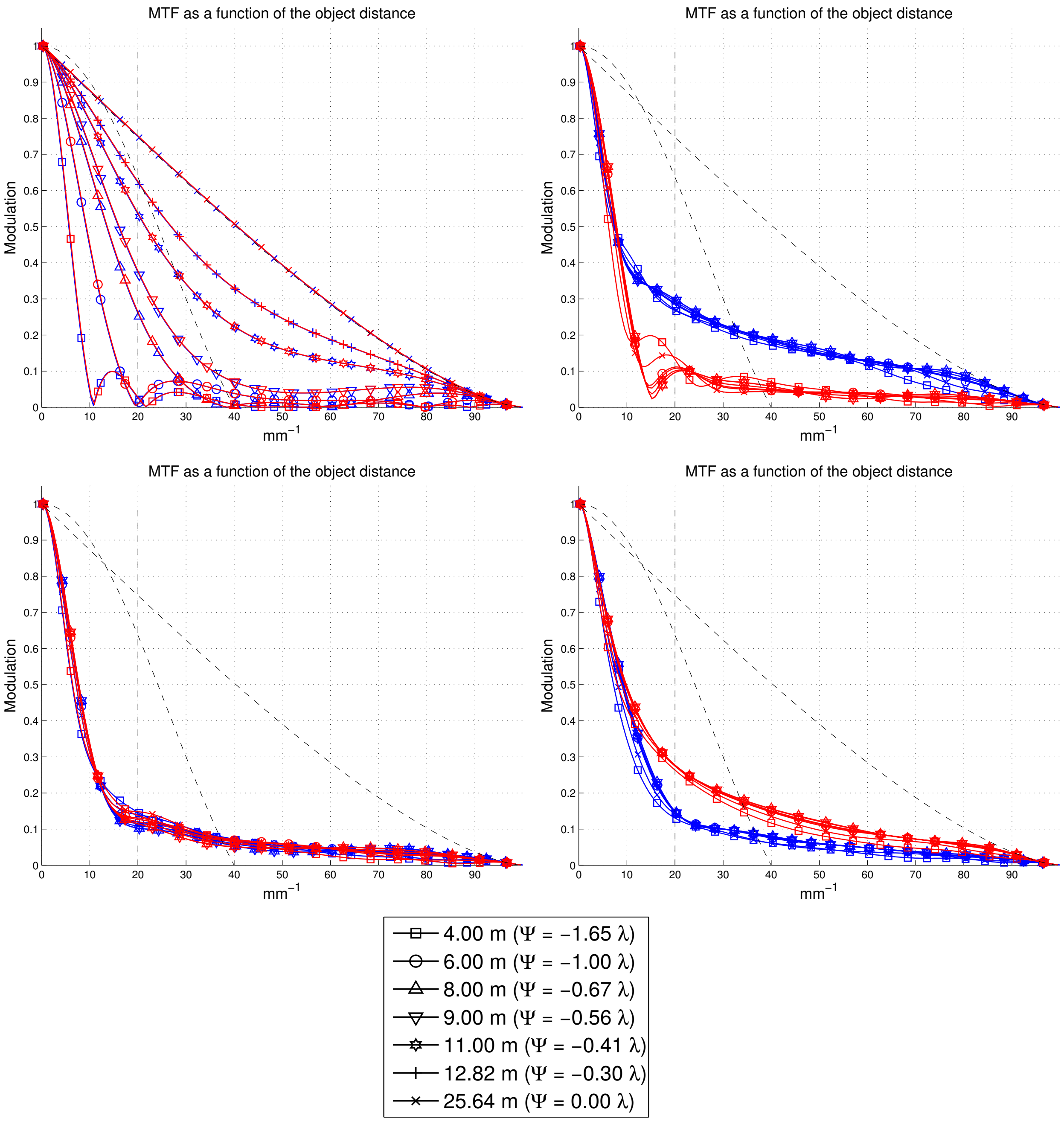


Figure 2.4: Comparison of the **MTFs** for a standard imager, common designs and the optimized phase mask. **Upper left corner** : conventional design (wide opened); **Upper right corner** : cubic phase mask design; **Lower left corner** : trefoil phase mask design; **Lower right corner** : optimized phase mask design. Blue MTFs are in X slice, red MTFs are in XY slice (at 45°).

parameter W_{SNR}. We share the selected value with the two other systems. The cubic phase mask design has a resulting noise gain of $4.5dB$, $1.5dB$ higher than the optimized design.

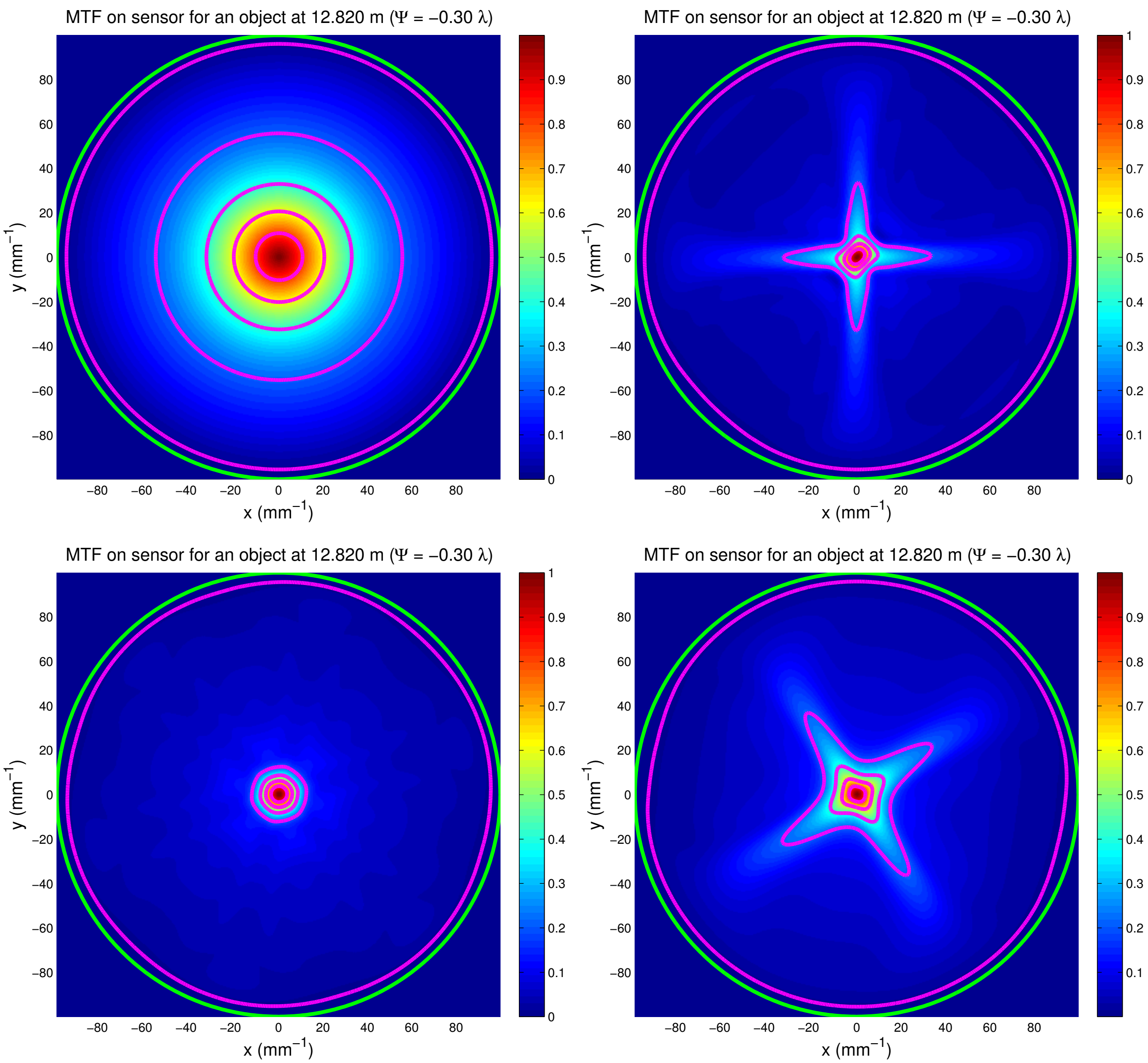


Figure 2.5: Comparison of the **MTFs** for a standard imager, common designs and the optimized phase mask. The object is at a distance of 12.8m ($\Psi = -0.30\lambda$). **Upper left corner** : conventional design (wide opened); **Upper right corner** : cubic phase mask design; **Lower left corner** : trefoil phase mask design; **Lower right corner** : optimized phase mask design. The isolines (purple) represents steps of 0.1 in MTF. The green contour corresponds to the optical cut-off frequency.

The trefoil phase mask design has a resulting noise gain of $4.8dB$, $1.8dB$ higher than the optimized design. We also test the image quality of these systems by simulating a Siemens Star pattern (Figure 2.10) and various natural scenes (Figures 2.12 and 2.14) covering the full OTF support (the virtual sensor is sampling up to the optical cut-off) for a SNR of

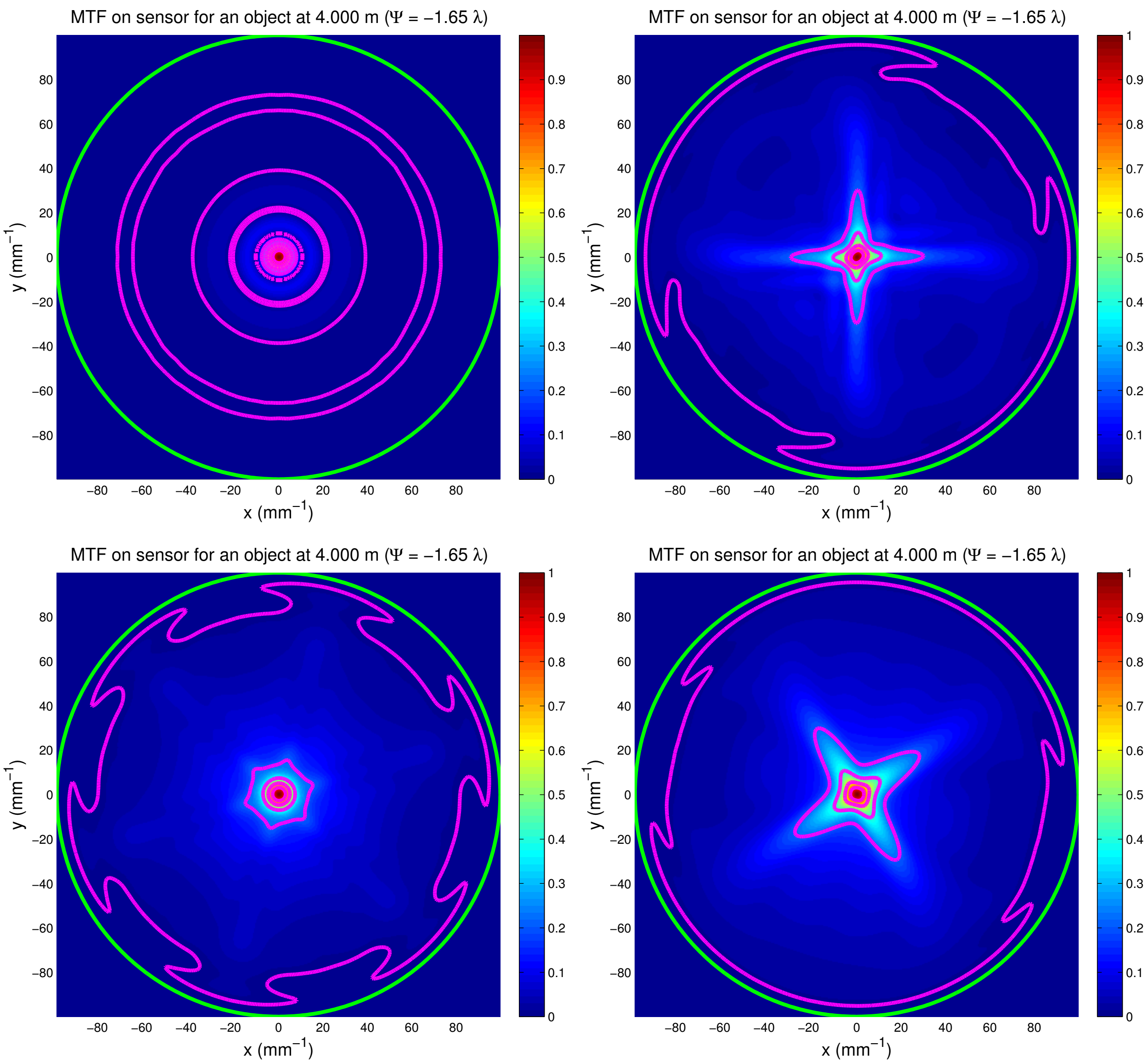


Figure 2.6: Comparison of the **MTFs** for a standard imager, common designs and the optimized phase mask. The object is at a distance of 4.0m ($\Psi = -1.65\lambda$). **Upper left corner** : conventional design (wide opened); **Upper right corner** : cubic phase mask design; **Lower left corner** : trefoil phase mask design; **Lower right corner** : optimized phase mask design. The isolines (purple) represents steps of 0.1 in MTF. The green contour corresponds to the optical cut-off frequency.

$25.0dB$ and with previous filters and reconstructions designs. The higher noise gain of cubic and trefoil-based designs over the optimized Zernike design is evident in these examples.

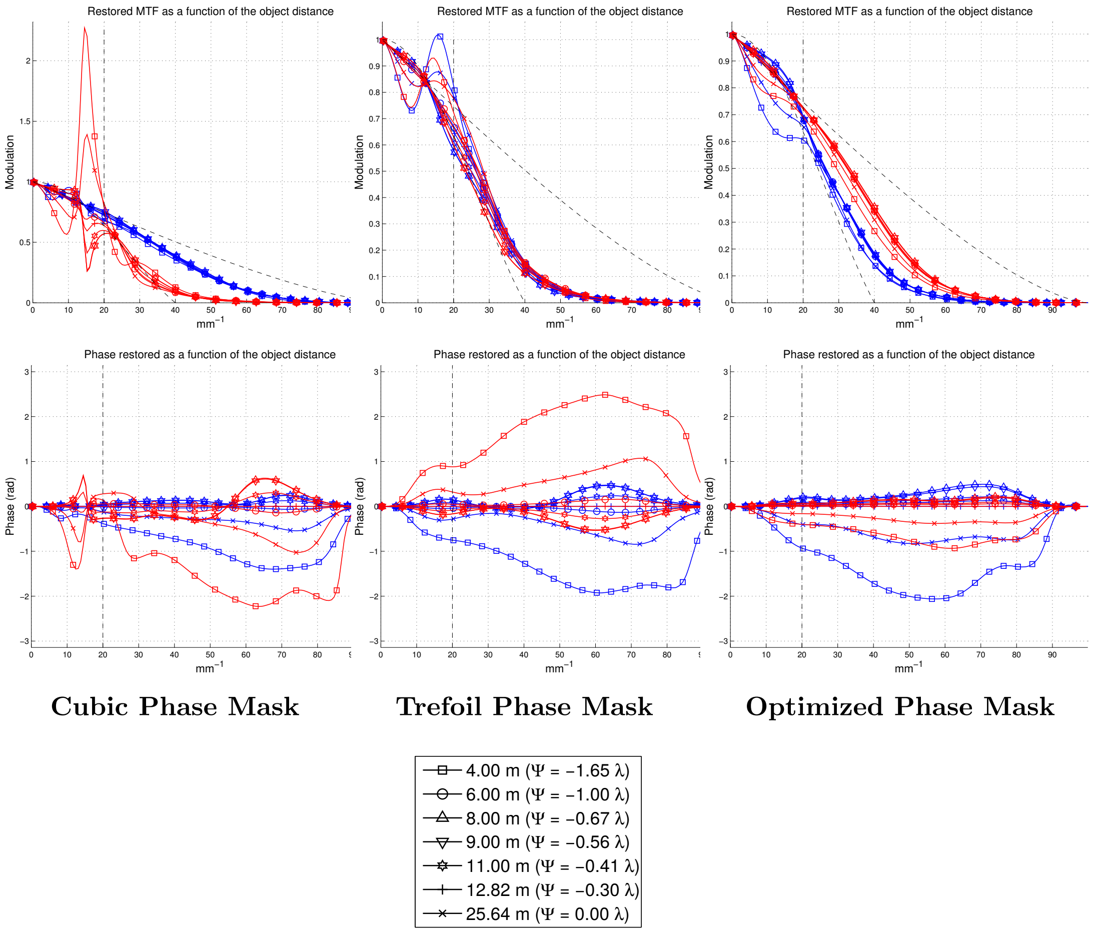
To highlight the artifacts generated by both the cubic and trefoil phase mask designs,

ID	Zernike Coefficient	(in unit of λ)	Expression
1	Defocus (n=2, m=0)	0.340	$(2\rho^2 - 1)$
4	Coma 1 (n=3, m=-1)	0.101	$(3\rho^3 - 2\rho)\sin(\varphi)$
5	Coma 2 (n=3, m=1)	0.561	$(3\rho^3 - 2\rho)\cos(\varphi)$
6	Trefoil 1 (n=3, m=-3)	-0.283	$(\rho^3)\sin(3\varphi)$
7	Trefoil 2 (n=3, m=3)	-0.321	$(\rho^3)\cos(3\varphi)$
13	(n=5, m=1)	0.092	$(10\rho^5 - 12\rho^3 + 3\rho)\cos(\varphi)$
14	(n=5, m=-1)	0.029	$(10\rho^5 - 12\rho^3 + 3\rho)\sin(\varphi)$
15	(n=5, m=3)	0.001	$(5\rho^5 - 4\rho^3)\cos(3\varphi)$
16	(n=5, m=-3)	-0.009	$(5\rho^5 - 4\rho^3)\sin(3\varphi)$
17	(n=5, m=5)	-0.024	$(\rho^5)\cos(5\varphi)$
18	(n=5, m=-5)	-0.020	$(\rho^5)\sin(5\varphi)$
26	(n=7, m=-1)	0.014	$(35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho)\sin(\varphi)$
27	(n=7, m=1)	0.065	$(35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho)\cos(\varphi)$
28	(n=7, m=-3)	-0.022	$(21\rho^7 - 30\rho^5 + 10\rho^3)\sin(3\varphi)$
29	(n=7, m=3)	-0.021	$(21\rho^7 - 30\rho^5 + 10\rho^3)\cos(3\varphi)$
30	(n=7, m=-5)	-0.033	$(7\rho^7 - 6\rho^5)\sin(5\varphi)$
31	(n=7, m=5)	0.003	$(7\rho^7 - 6\rho^5)\cos(5\varphi)$
32	(n=7, m=-7)	-0.027	$(\rho^7)\sin(7\varphi)$
33	(n=7, m=7)	0.001	$(\rho^7)\cos(7\varphi)$

Table 2.4: Optimized phase mask coefficients.

due to the over or undercompensation of the System Transfer Function at particular distances as shown in Figures 2.7, 2.8 and 2.9, we remove the noise randomness from the simulations while maintaining all other parameters constant (including Wiener filter prior parameters constant). These synthetic results are shown in Figures 2.11, 2.13 and 2.15. The optimized design shows again a superior and sustained image quality throughout the range of distances tested. Note finally that all of the EDoF systems tends to deliver a slightly softer image at both edges of the specification range, *e.g.* 4.0m and infinity in the current context. This is due to the change of focus provoked by the only parabola term contained in the phase masks expression, which centers the imaging performances in the middle of the defocus range of interest, *e.g.* at about 12.8m here.

Although the coefficients corresponding to the polynomials of the 5th and 7th orders are small in magnitude compared to those of the 3rd order and the defocus term, their contribution to the overall system performance is significant. As shown in Figure 2.18,



Cubic Phase Mask **Trefoil Phase Mask** **Optimized Phase Mask**

Figure 2.7: Comparison of the **STFs** for major designs. **Top row** : restored amplitude modulation; **Bottom row** : restored phase. Blue MTFs are in X slice, red MTFs are in XY slice (at 45°).

if we truncate these higher order components from the optimized mask the MTF is not kept constant anymore through the distance range. After reconstruction, this distortion leads to a large amplitude variation of the transfer function at mid- spatial frequencies. Thus, this truncated mask design introduces artifacts in the restored images similar to those encountered for the cubic phase mask.

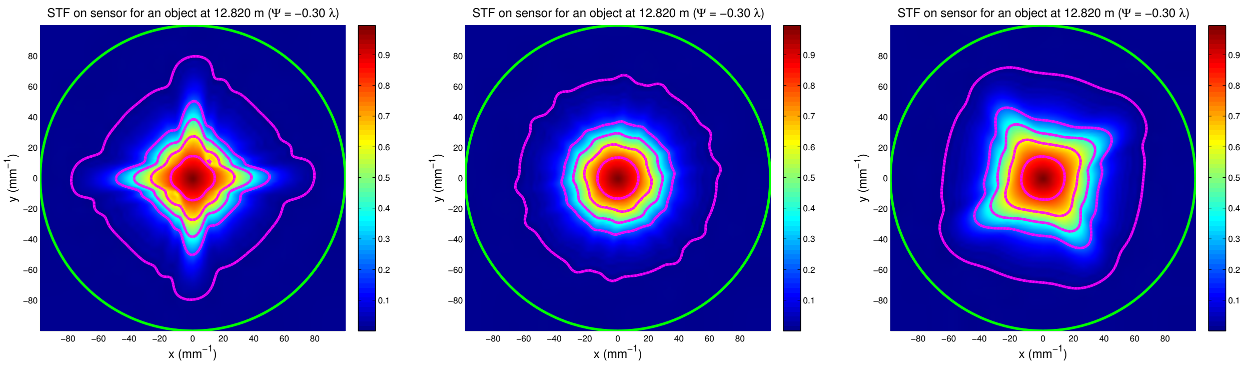


Figure 2.8: Comparison of the **STFs** for major designs for an object at a distance of 12.8m ($\Psi = -0.30\lambda$). **From left to right** : cubic phase mask design, trefoil phase mask design, optimized phase mask design. The isolines (purple) represents steps of 0.1 in MTF. The green contour corresponds to the optical cut-off frequency.

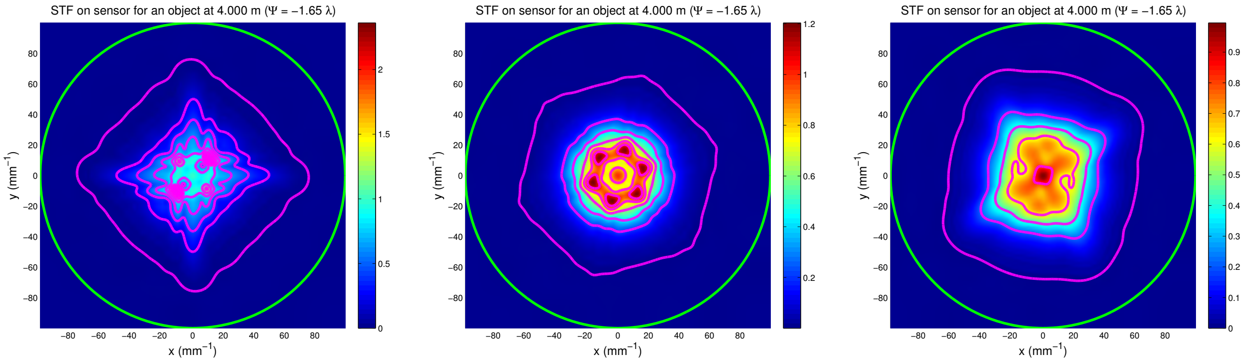


Figure 2.9: Comparison of the **STFs** for major designs for an object at a distance of 4.0m ($\Psi = -1.65\lambda$). **From left to right** : cubic phase mask design, trefoil phase mask design, optimized phase mask design. The isolines (purple) represents steps of 0.1 in MTF. The green contour corresponds to the optical cut-off frequency.

2.4 Experimental results

To validate the performance of the optimized system, we build the prescribed Zernike optimized phase mask out of a $\varnothing 25mm$ germanium flat mask (with refractive index $n \approx 4.0$) with a freeform diamond turning CNC machine and test that the surface is meeting the design specifications via interferometric optical testing. To compare the performances of the computational and conventional designs, we use two identical thermal cameras with each an uncooled microbolometers array of size 384×288 pixels and $25\mu m$ pitch matching the

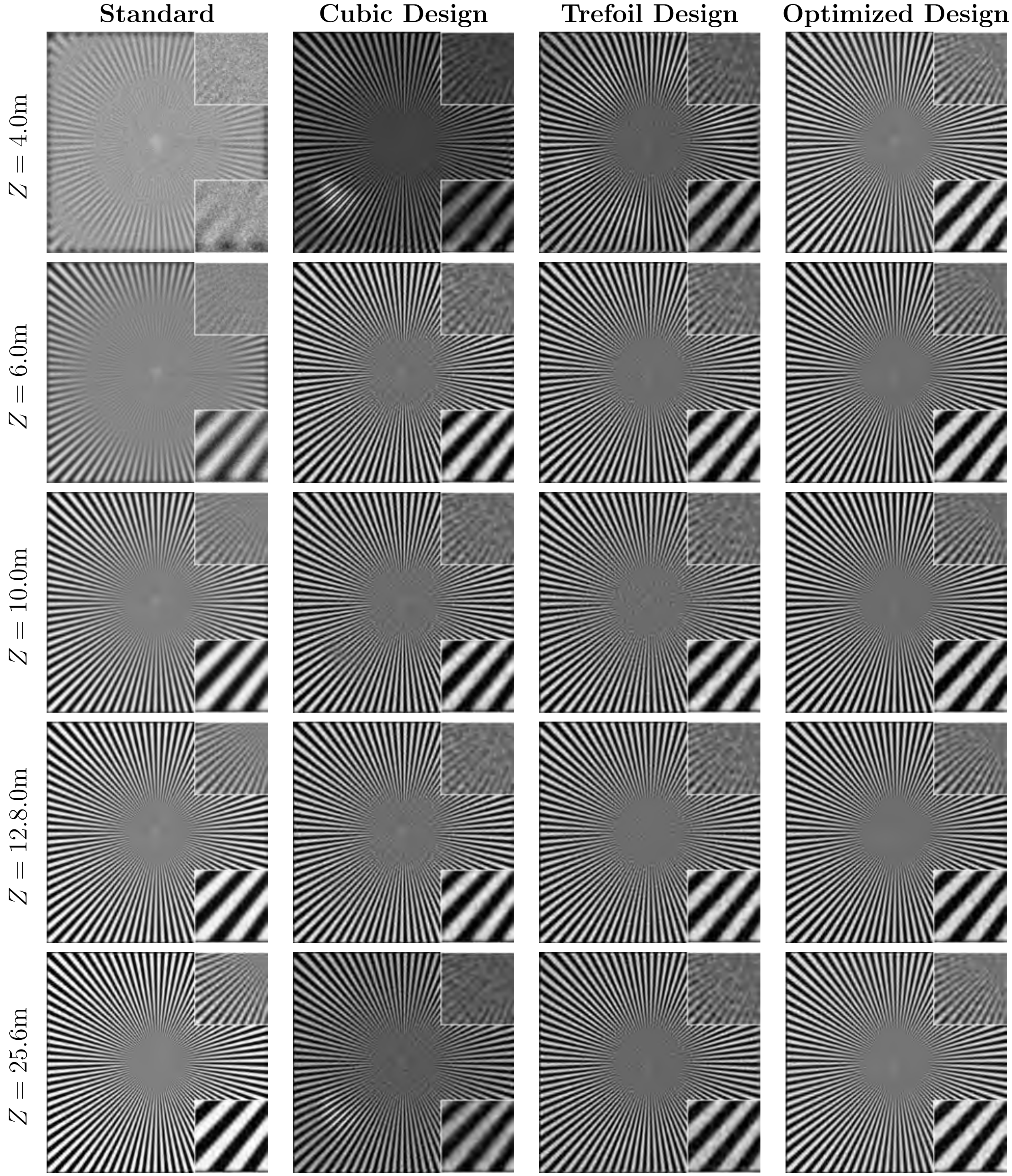


Figure 2.10: Simulation of images over the full OTF support after reconstruction, for a SNR of 25dB. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

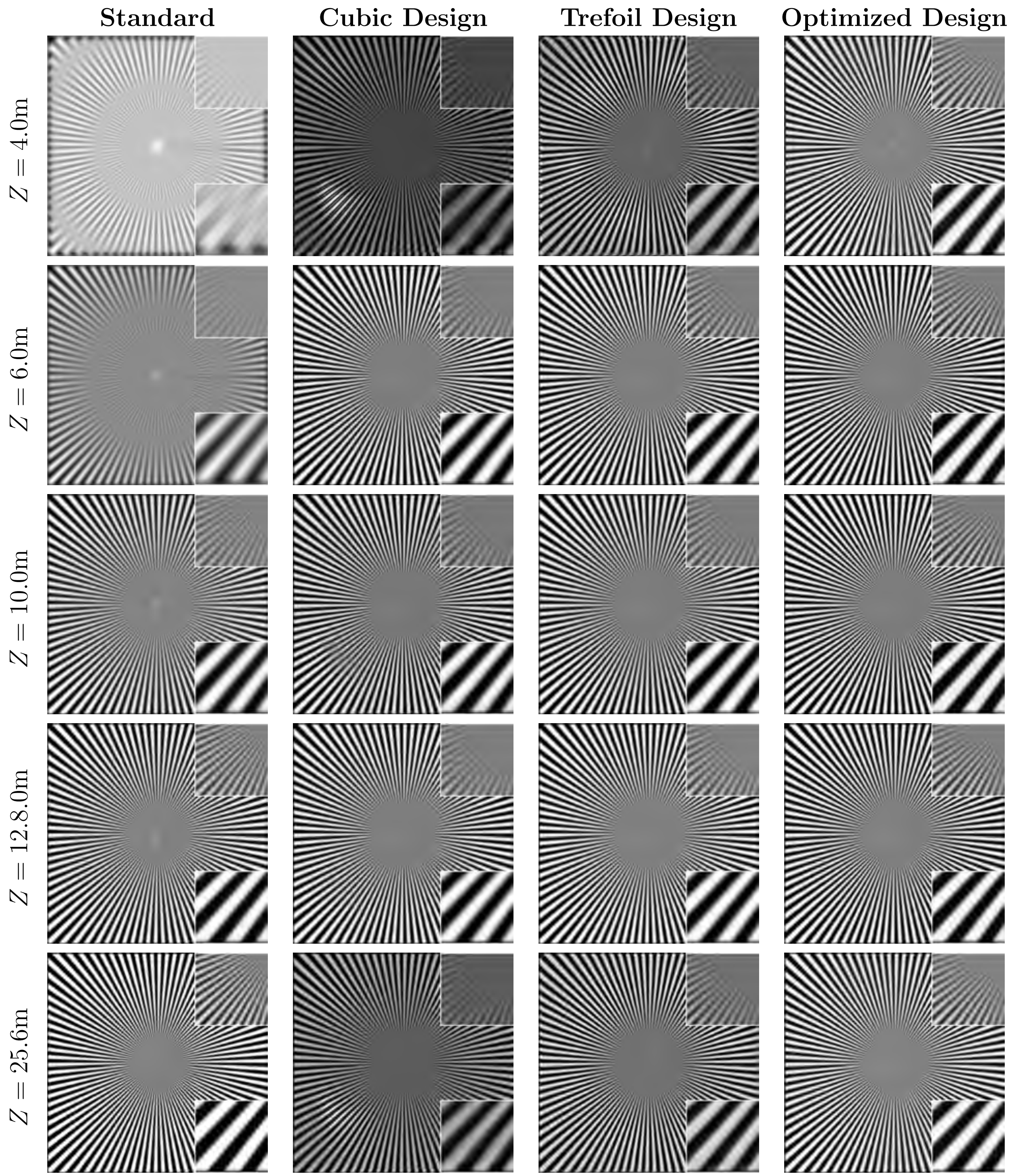


Figure 2.11: Simulation of images over the full OTF support after reconstruction, excluding noise. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

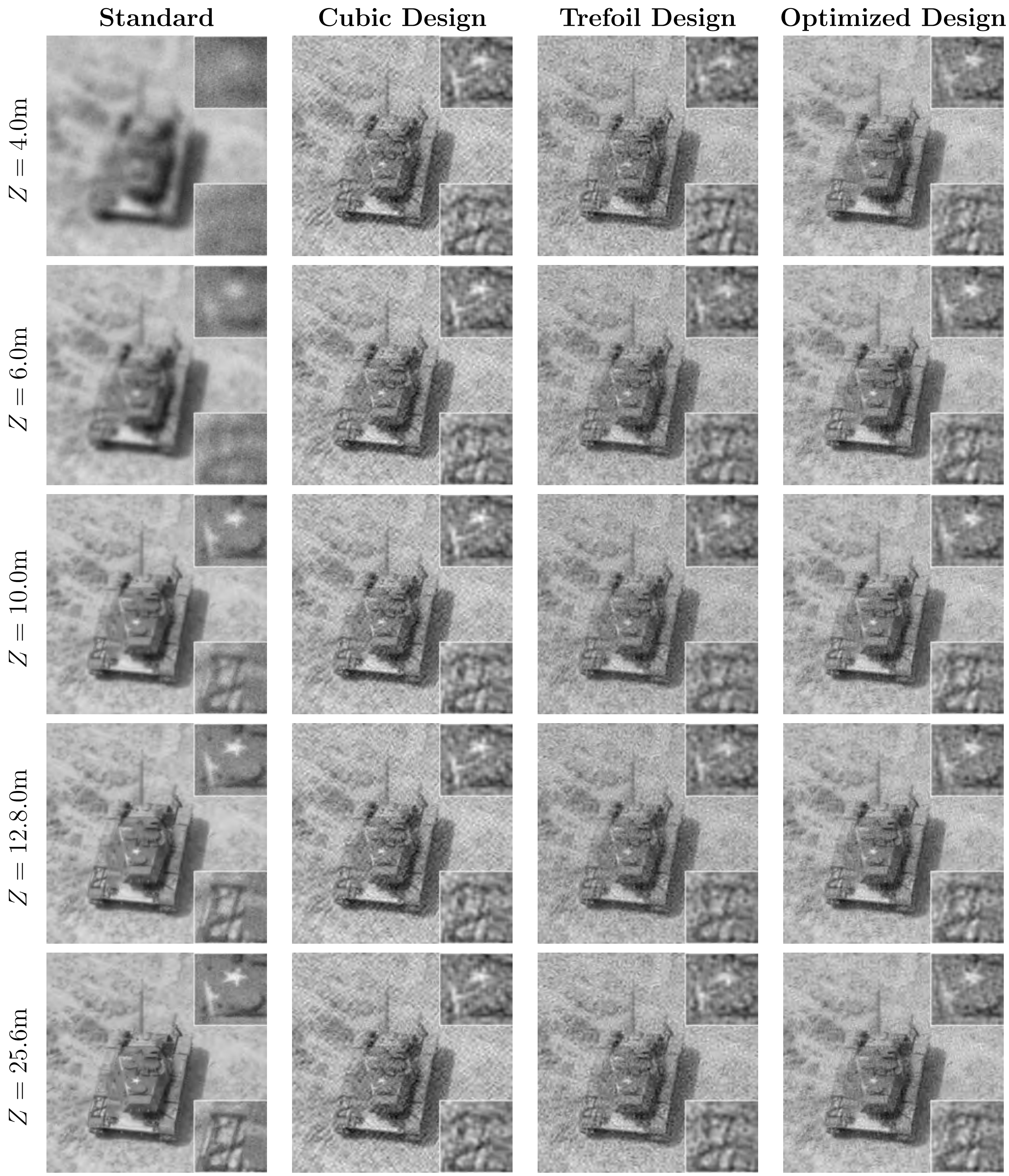


Figure 2.12: Simulation of images over the full OTF support after reconstruction, for a SNR of 25dB. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

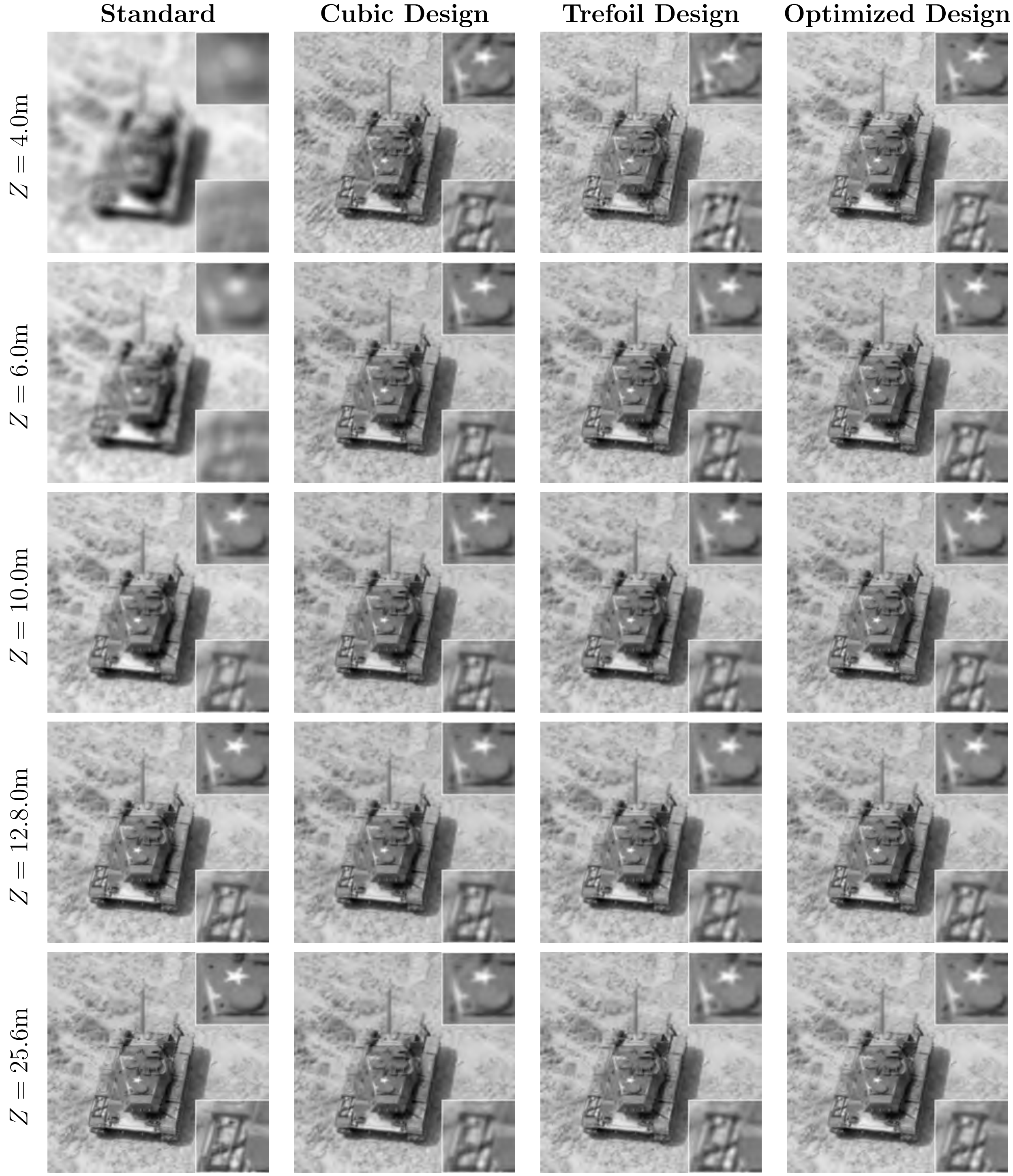


Figure 2.13: Simulation of images over the full OTF support after reconstruction, excluding noise. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

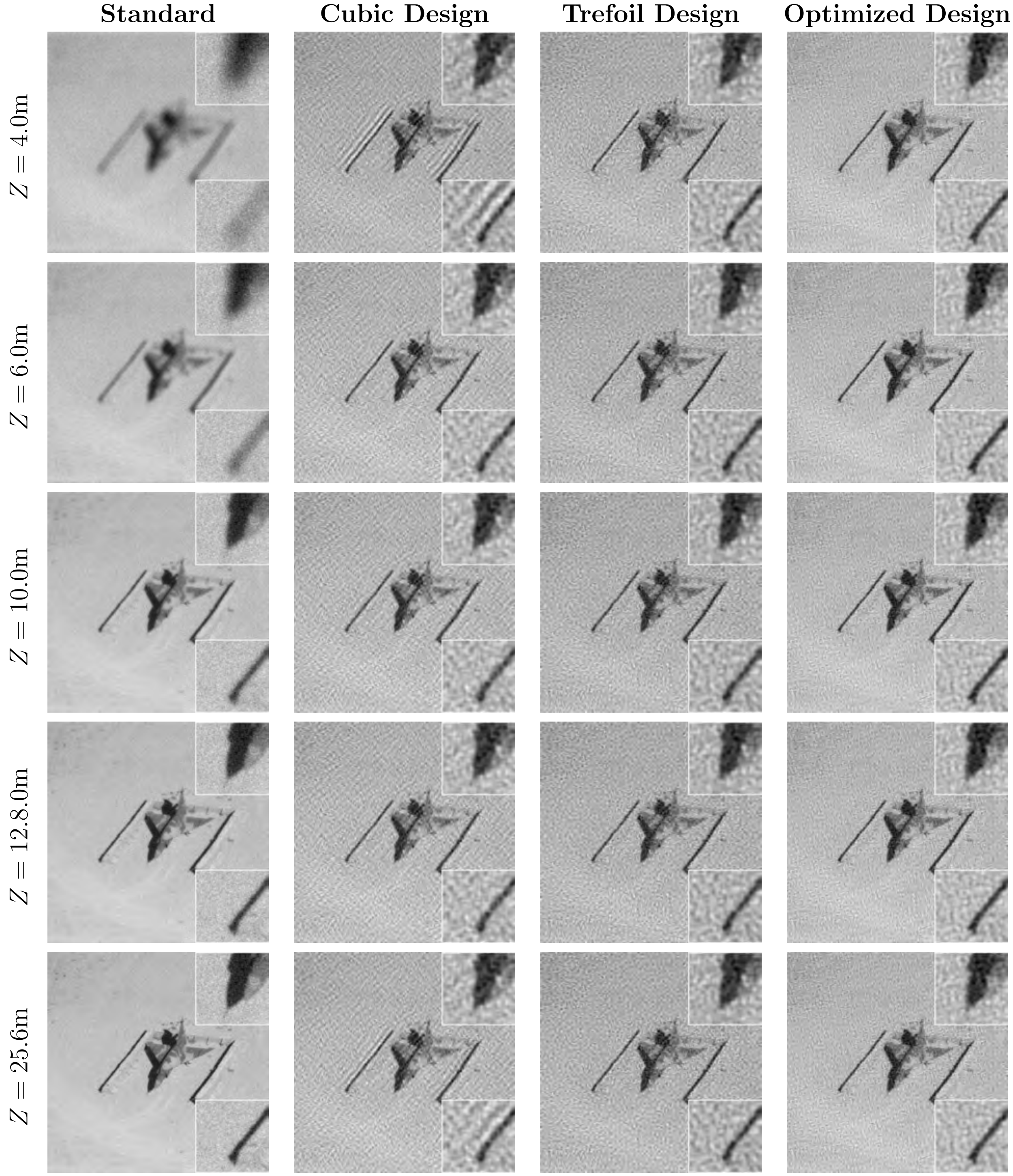


Figure 2.14: Simulation of images over the full OTF support after reconstruction, for a SNR of 25dB. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

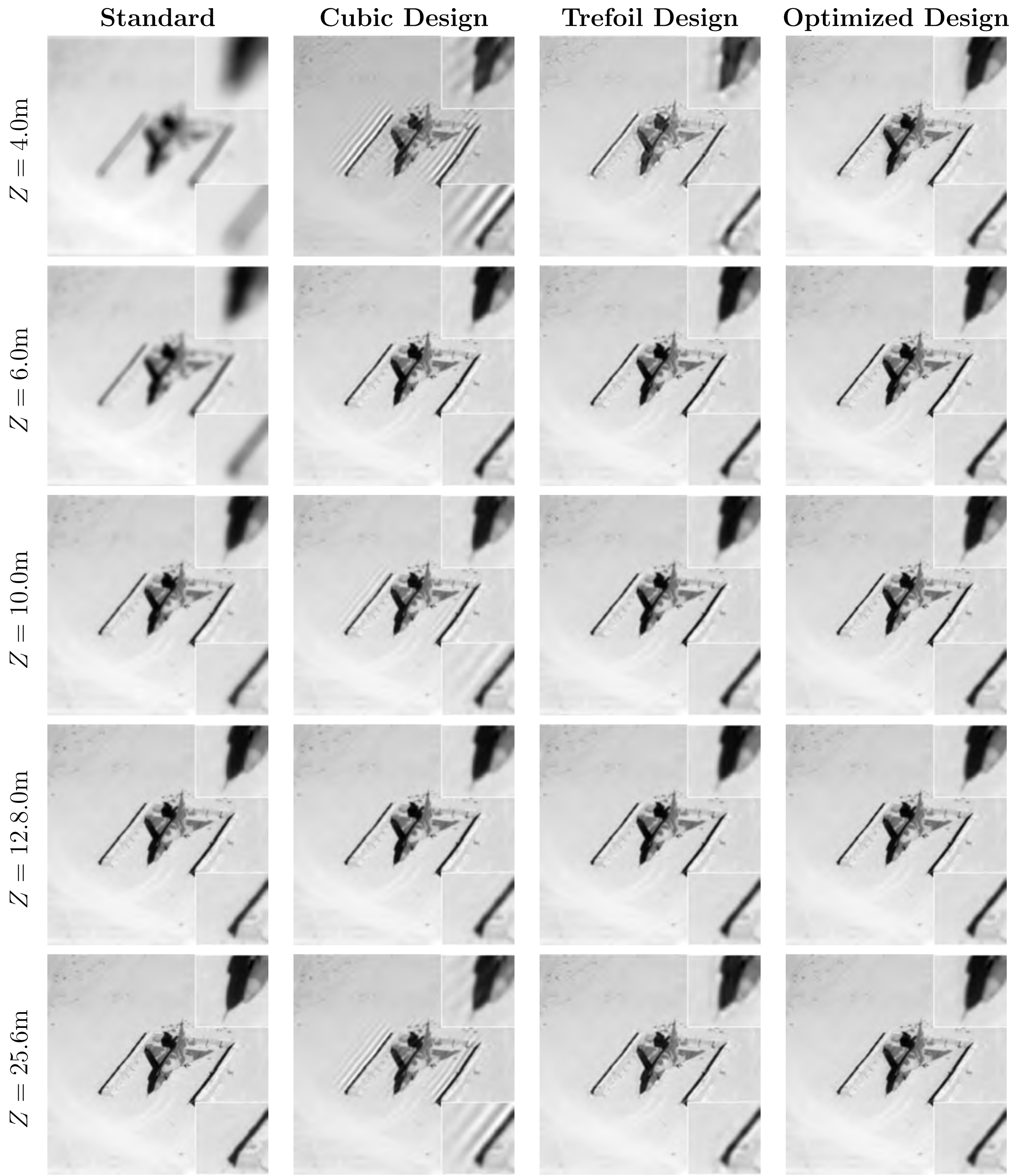


Figure 2.15: Simulation of images over the full OTF support after reconstruction, excluding noise. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

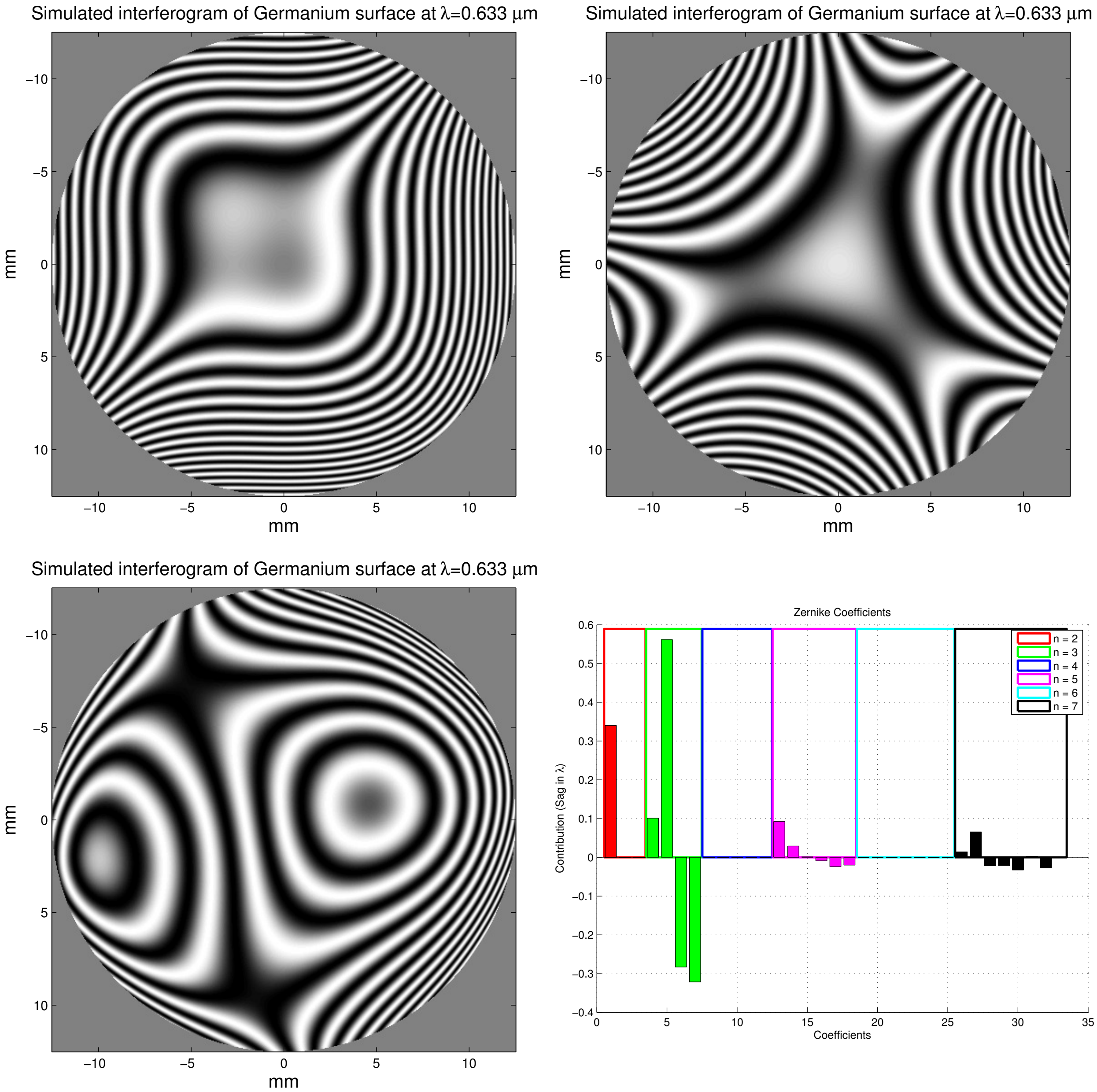


Figure 2.16: Simulated interferograms of the surfaces at $\lambda = 0.633\mu m$. **Upper left corner** cubic phase mask design; **Upper right corner** : trefoil phase mask design; **Lower left corner** : optimized phase mask design; **Lower right corner** : Zernike Coefficients bar chart for the optimized Zernike surface.

specifications of Table 2.1. These cameras especially operate with a limited SNR budget of about $20dB$. They also use two identical $25mm$ F/1 germanium lenses consisting of two aspherical elements and which are diffraction-limited for the in-focus plane. We secure the germanium phase mask at the entrance aperture of one of the camera and will apply the

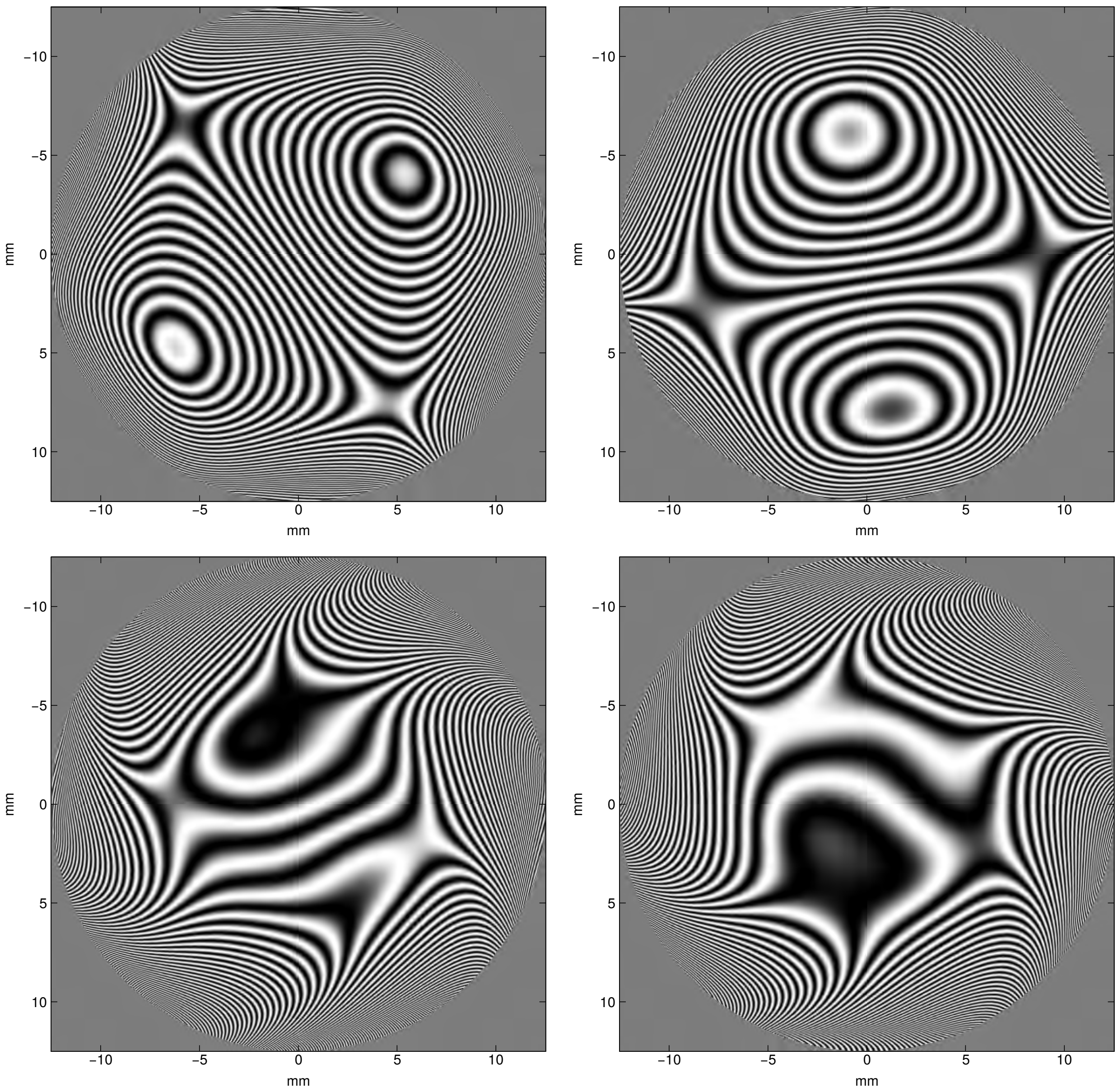


Figure 2.17: Other surfaces returned by the optimization process for the same operating characteristics (simulated interferograms at $\lambda = 0.633\mu m$). Most of the results can be sorted in one of the two largest sets depending if they are based primarily upon coma (**top row**) rather than trefoil (**bottom row**) Zernike coefficients.

linear digital reconstruction filter described previously to its output images.

To test the performance of both the conventional and the computational system, we image a calibrated blackbody source with its surface temperature set to $37°C$. We place different bar chart masks in front of the source to analyze the modulation transfer functions of

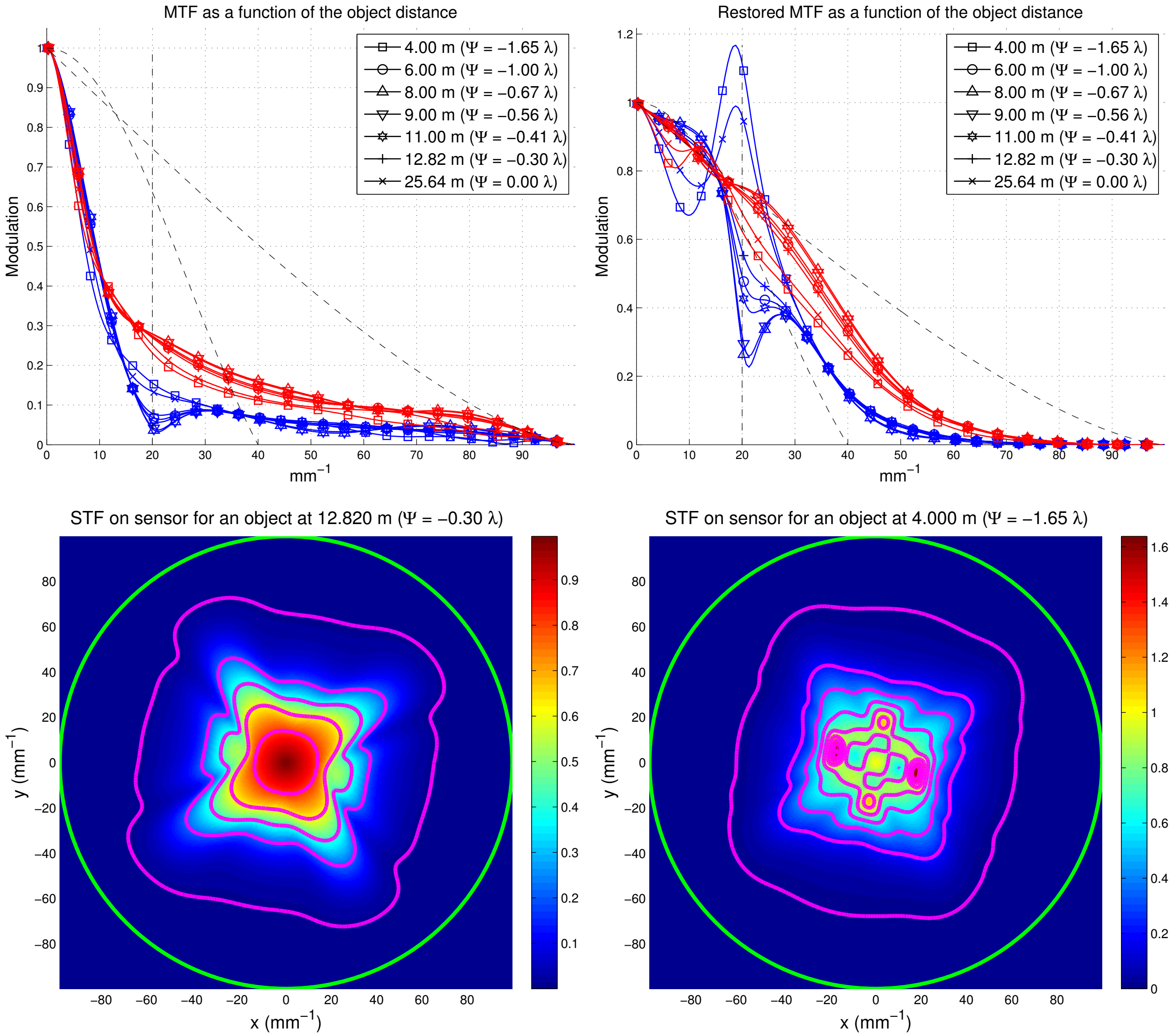


Figure 2.18: Modulation transfer functions after truncation of the 5th and 7th orders Zernike polynomials from the optimized surface. **Upper left corner** : optical MTF; **Upper right corner** : STF; **Lower left corner** : 2D STF at a distance of 12.8m ($\Psi = -0.30\lambda$); **Lower right corner** : 2D STF at a distance of 4.0m ($\Psi = -1.65\lambda$).

the two systems. In Figure 2.19 we show several of the acquired images (after reconstruction in the case of the computational imager) for distances ranging from 3.0m (which is outside of the optimization range) to 6.0m. The superior sharpness of the computational imager is evident at these close ranges by looking at the modulation profile across the images, as illustrated in Figure 2.20. We especially note that the image quality delivered by the optimized phase mask decay gracefully for low spatial resolutions as shown for a distance of

$3.0m$ and a spatial frequency $\eta \leq 14mm^{-1}$. Finally, Figures 2.21 and 2.22 show respectively an indoor scene with a person standing at 4.0m of both cameras apertures for the first and an outdoor scene composed of vegetation in the foreground, at about $5m$, and buildings in the background, at about $350m$.

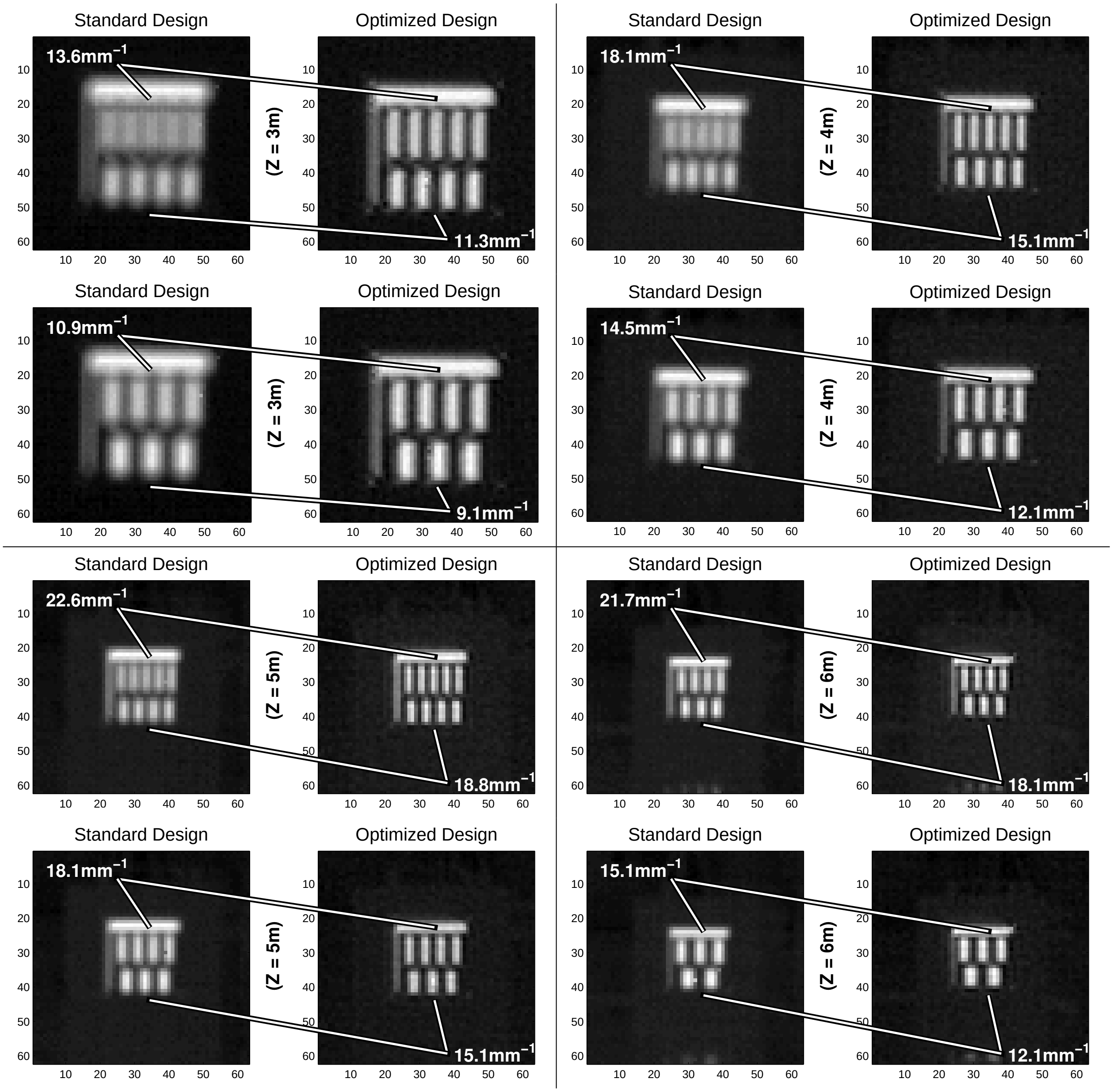


Figure 2.19: Experimental results of several bar charts patterns in front of a blackbody source at 37°C.

2.5 Conclusion

We presented a novel optimization framework for the design of computational EDoF imagers which can deliver a sharp image on a broader range of object distances than a conventional system. Contrary to the constraining optical image quality criterions related to the aberrated wavefronts root mean square error by Strehl and others [10, 11, 12, 13], the aberration induced by a phase mask in this computational imaging system effectively neutralizes the effect of defocus. The resulting soft optical image can then be restored to diffraction-limited performance by a digital processing algorithm. In this work, we particularly focused on maximizing the image quality delivered by the system and, for this, we proposed an optimization framework which relies on (*1*) the use of the orthonormal Zernike polynomials to model the phase mask surface and (*2*) a metric assessing the overall modulation transfer function of the system, that is after reconstruction by the optimal linear filter. We used the simulated annealing algorithm to optimize the parameters of the optical phase mask to meet specifications in the case of a 25mm F/1 thermal imager. Interestingly, we have seen that only the odd orders Zernike polynomials (excluding the main defocus term) provide depth of field insensitivity and that, although the coefficients of the 5th and 7th orders polynomials are small in magnitude, they contribute significantly to keeping the MTF constant across the range of distances. We simulated the resulting image quality of one of the optimized phase mask and showed its superior performance over the (similarly optimized) cubic and trefoil designs as it generates less artifact and has a lower noise gain, by more than $1.6dB$. Finally, we built, tested and validated the performance of the optimized design against a conventional imager with real-world scenes.

In the future, we can port this computational imaging approach from the extension of the depth-of-field to the extension of the depth-of-focus which is particularly valuable for small form-factor imaging modules suffering from a large field curvature [27]. In the context of this work, we also only focused on a long-wave infrared application, for which the dispersion of the refractive materials is limited because the spectral range between 8μm and 12μm is relatively narrow when compared to the center wavelength of 10μm. Therefore, we were able to approximate all the radiation to a monochromatic profile. However, for visible applications, *i.e.* for λ from 400nm to 650nm, this approximation is not valid. We can note the interesting approach chosen by Guichard *et al.* in [14, 17] as they use axial chromatic aberrations to extend the depth of field by transporting sharp texture information across the three RGB sensor channels. This approach could be employed jointly with the optical phase mask design discussed here in order to further increase the depth range and adapt to broad spectral range.

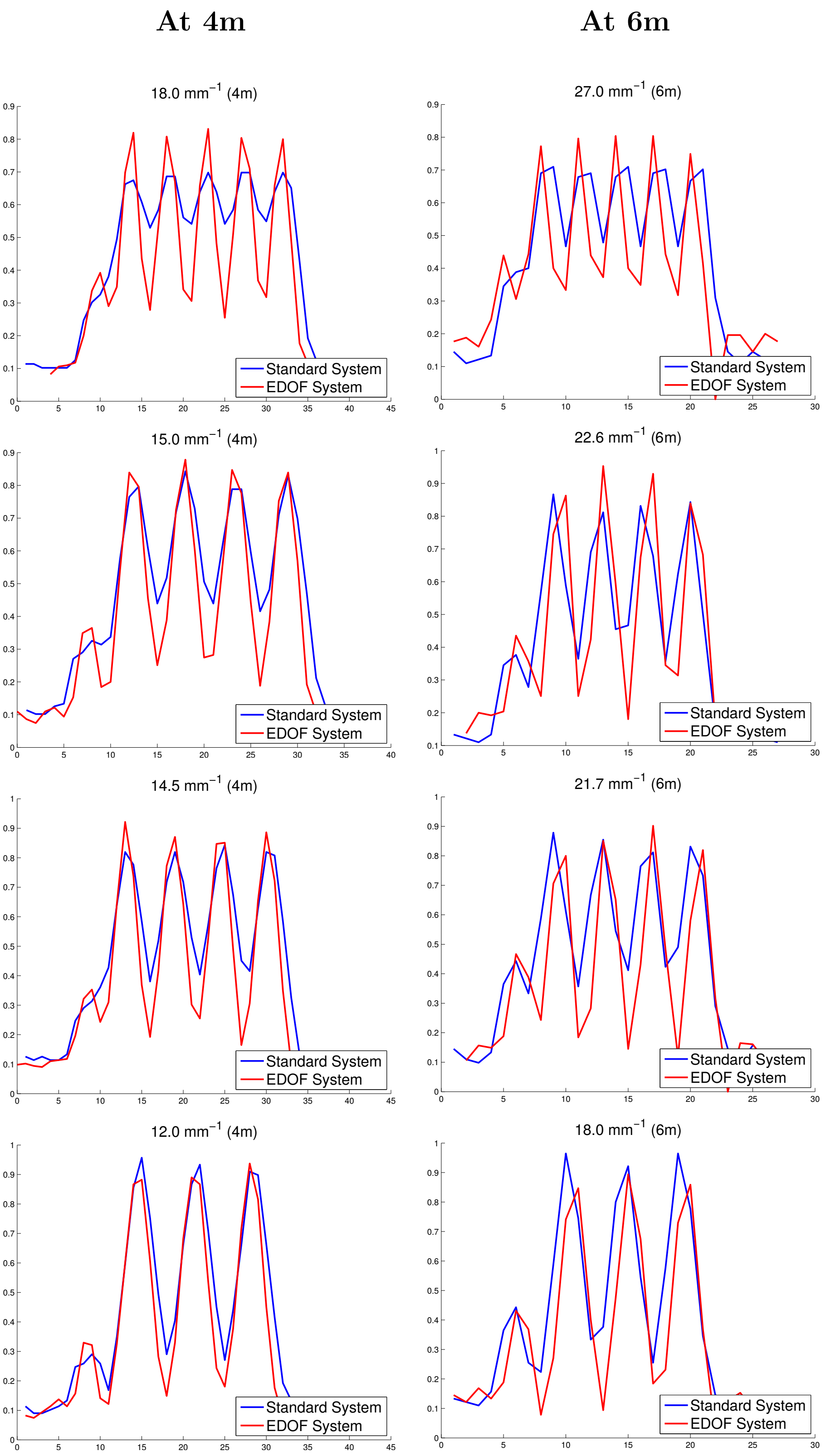


Figure 2.20: Cross-section profiles of the previous experimental data showing the difference in modulation between the Standard Design and the Optimized Design.

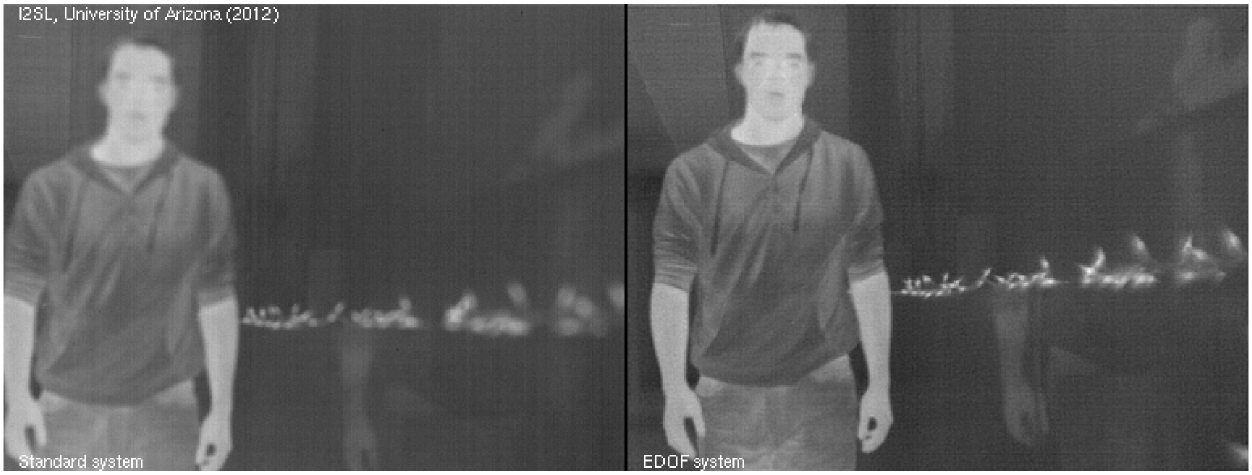


Figure 2.21: Side by side comparison of a natural scene. On the left is the image produced by the standard system (a FLIR camera) and on the right, by the EDoF system (an identical FLIR camera, with the phase mask mounted and after the digital reconstruction of the stream). For both side, the cameras are focused at 25.0m and the person is standing at approximately 4.0m. Note that the closest light bulbs on the right side are at a distance of 2.0m.

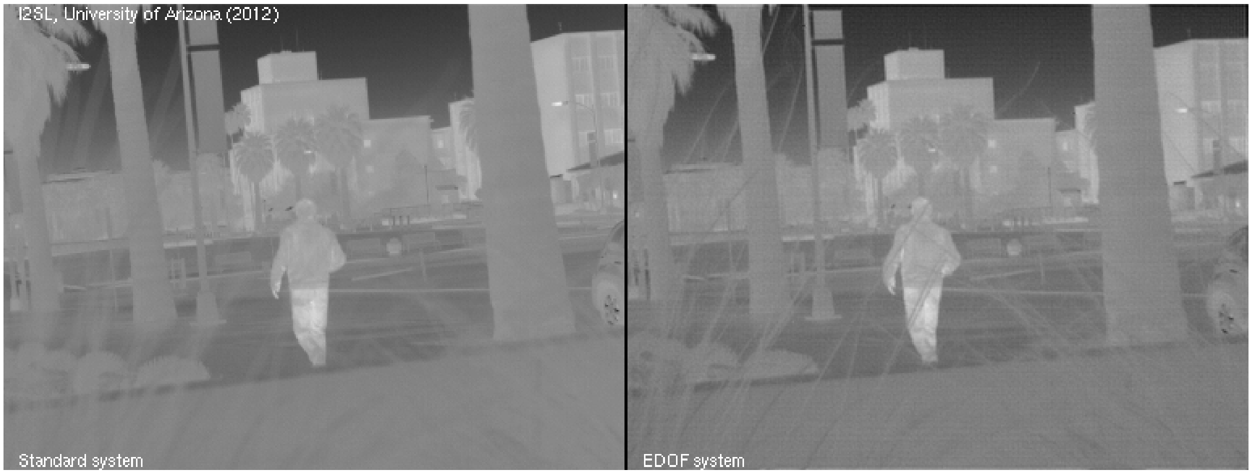


Figure 2.22: Side by side comparison of a natural scene. On the left is the image produced by the standard system and on the right by the EDoF system. Note that the foreground vegetation is at approximately 5m of the two systems apertures and the buildings in the background are at a distance of approximately 350m.

Chapter 3

Scalable Compressive Imaging System Design and Implementation

Compressive Imaging is a new design paradigm which notably allows to free the sensor array from Nyquist's sampling requirement. While this technique can be implemented to effectively acquire large images from a single photo-detector, it necessitates to perform a computationally costly and slow reconstruction. In this chapter, we present a block-wise parallel acquisition architecture and reconstruction algorithm to record and process high-resolution images in real-time. We implement it as a programmable testbed which allows us to verify the superior performance of previously developed Information Optimal projection patterns versus commonly used random projections.

3.1 Introduction

While all conventional cameras sample densely the field of view, compressed sensing (CS) exploits the sparsity of natural scenes and allow for a complete image reconstruction from fewer measurements [28]. This is implemented, in the context of incoherent illumination, by a linear measurement model where the scene signal-vector is modulated by a series of patterns before being integrated onto a few intensity detectors. With this strategy come several gains over traditional devices. First, the number of detectors needed for high-resolution imaging is considerably less for this architecture than a standard focal plane array (FPA); this reduction in sensor complexity is welcomed for cameras working in spectral domains where the per-pixel cost is prohibitive, as it is the case, for instance, in short-, mid- and long-wave infrared (respectively noted SWIR, MWIR and LWIR) as well as terahertz. Secondly as the modulation is performed in the optical domain, it escapes the noise corruption inherent to the detection of the light signal. Thus, the compressive imager benefit from the multiplexing gain applied to the signal and can therefore operate in a low signal-to-noise ratio (SNR) regime. Finally, the high-resolution signal is acquired directly into a compressed representation which translates into a significant reduction of the sensor readout bandwidth, reducing both its thermal envelope (and decreasing the associated thermal noise) and its power consumption. In addition, these novel architectures do not require the digital compression of the high-dimensional image stream at high frame-rate which is otherwise hindering the autonomy of embedded systems.

The single pixel camera (SPC) [29, 31] was the first implementation of an optical compressive imaging architecture. It utilizes a single photo-detector to integrate the spatially-modulated full field of view leading to an extreme reduction in the complexity of the detector from the large arrays used by conventional imaging architectures. However, this particular architecture comes with two downsides. First, for a constant compression-ratio it requires to multiply the number of modulation patterns used to adapt for higher space-bandwidth products (*i.e.* resolution × field of view). Second, the associated reconstruction complexity for all specialized algorithms is at least in the order of $\mathcal{O}(N^2)$ where N is the dimension of the scene vector. Consequently, the runtimes per frame for these iterative reconstruction processes are in the order of seconds to minutes and forbid the use of the SPC in a real-time environment. Therefore, just as traditional imaging systems, the SPC does not scale well to high space-bandwidth products.

To overcome these limitations, we have proposed in [40, 41] a scalable CS Imaging

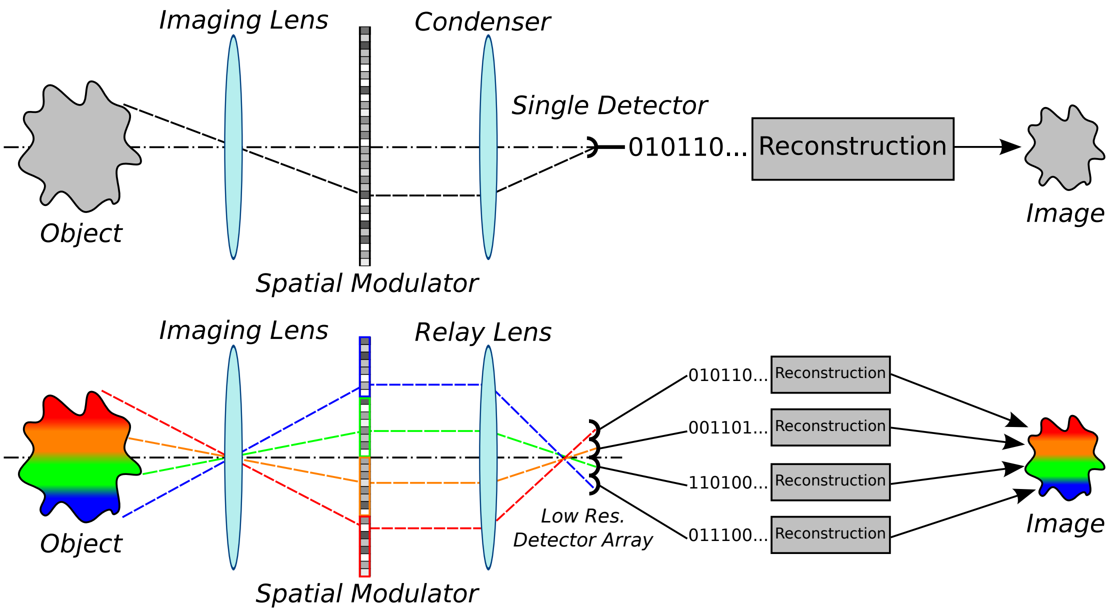


Figure 3.1: **Top** : Single Pixel Camera architecture, the full field of view is conjugated to the SLM plane where its is modulated and integrated, from there, onto a single detector. The patterns to be used for the projection are consecutively displayed. **Bottom** : Scalable Compressive Imager, the field of view is conjugated to the spatial modulator which is divided into blocks. Each block is mapped to a single detector and reconstruction unit and all can operate in parallel.

architecture relying on a block-wise decomposition of the field of view. Each block is independently projected against a unique series of patterns and the modulated irradiance is integrated onto a low resolution sensor. Each of the pixel in the focal plane array (FPA) thus collects the modulated flux from exactly one block. This parallel architecture can be represented simply as the concatenation of smaller independent SPCs for both acquisition and reconstruction, as illustrated in Figure 3.1. It is also thereby positioned within a spectrum of CS imager configurations ranging from conventional imagers, which are using large FPAs only, to the SPCs, which are using a large spatial modulator and combine all the output flux onto a single photo-detector. They can thus be parametrized by a *block size* defined as the number of modulation elements per detector element. Hence, for a traditional imager architecture, this parameter is virtually equal to one, and it is in the order of several millions for the SPC (assuming that a megapixel-scale modulator is employed). However, by appropriately tuning this parameter we can select a configuration performing optimally for a given spectral band, sensor technology and cost. Finally, besides the acquisition parameters, this parallel measurement architecture offers the possibility of employing a block-wise parallel reconstruction algorithm to reconstruct images in near real-time.

Note that similar compressive imager design have been implemented since : in [42], the authors have presented a one dimensional tiling version of the architecture with a linear sensor array and fast acquisition rate which is limited by constraints similar to the SPC for higher resolutions along the direction orthogonal the 1D FPA. The article [43] demonstrates a similar 2D tessellation approach in the SWIR band, but employing a global reconstruction algorithm. Finally, [44] focuses on the calibration aspect of such devices, particularly in measuring the distortion between ideal and experimental sensing matrix but without acquiring compressed measurements per CS definition.

In this chapter we describe the design, technical characteristics and implementation considerations of such a scalable compressive sensor as a fully programmable testbed. First, in section 3.2 we describe the components of this optical architecture implementation. We then present the measurement model in section 3.3, followed by the scalable block-wise PLE-MMSE reconstruction algorithm in section 3.4. We use those to assess the performance of Information-Optimal projections compared to random projections via a simulation study, in section 3.5, as well as experimental measurements with our prototype implementation, in section 3.6, for which we describe a fully automated calibration method that is critical to the system performance. Finally, we present our conclusions in section 3.7.

3.2 Scalable Optical Architecture

The structure of the compressive imager shown in Figure 3.2 can be decomposed into two sections : the first produces an optical image of the scene analogously to the objective of a traditional camera while the second modulates the image formed and maps it onto a low resolution sensor. We call the first part the *imaging arm*, which we designed from an off-the-shelf objective followed by an optical relay to increase the back-focal distance, both visible in Figure 3.3. It operates at $F/8$ and covers a 40° degrees full-field-of-view. Instead of

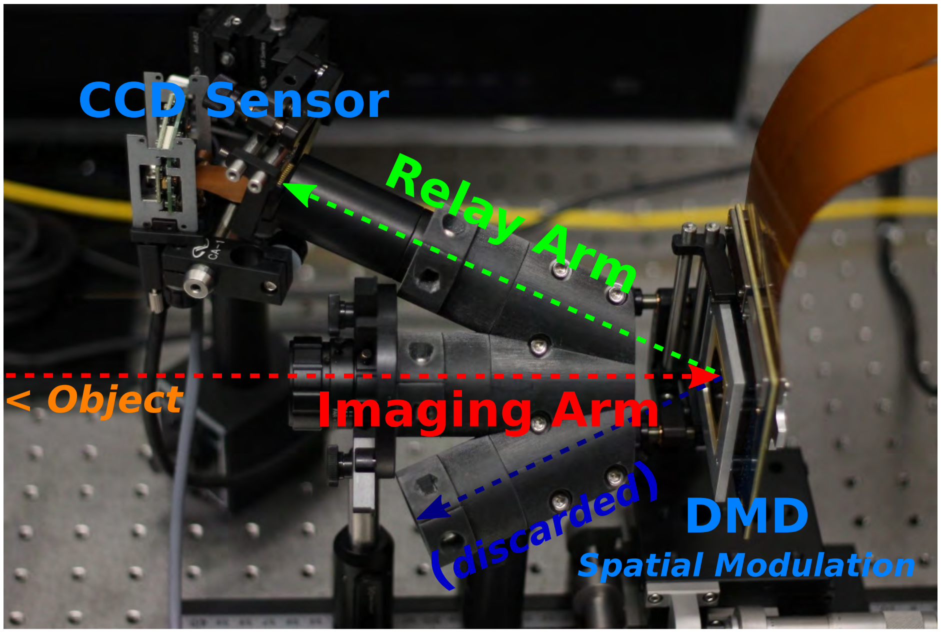


Figure 3.2: Picture of the Scalable CS Prototype and schematic overlay.

recording the image with a sensor as it is the case in the conventional architecture, a passive spatial modulator is placed in the image plane. This element will ultimately define the spatial resolution of the compressive imager : for this prototype, we are using a 1920 × 1080 Digital MicroMirror Device (DMD). This affordable component is commonly found in video projectors and consists in an array of bistable MEMS mirrors, 10.8μm in size. Every single of the two millions facets can be electronically controlled to tilt at an angle of either plus or minus twelve degrees around the array main diagonal, and effectively steers the normally incident beam coming from the imaging arm to either plus or minus twenty-four degrees. To achieve gray-scale modulation, these mirrors are rapidly switched between the two states (*on* or *off*) during the exposure : the ratio of the time spent in the *on*-state over the total duration determines the effective intensity reflection coefficient. All the micromirrors are synchronized to the same electronic control clock signal oscillating at 76.8KHz and can thus perform 8 bits quantized gray-scale modulation at a maximum refresh-rate of 300Hz. Therefore, the minimum possible exposure duration for gray-scale projection patterns in this compressive imager is about 3ms.

The second arm or *relay arm* is consequently at an angle of 24° from the DMD normal, around its main diagonal, and has a F-Number matched to that of the imaging arm. It maps the modulated image from the DMD plane to the surface of a low resolution sensor array and consists of custom designed optics to minimize aberrations and distortions affecting this critical transformation. For the sensor, we use a 1/3" 640×480 high-speed mono-CCD with a pixel size of 7.4μm and 12-bits ADC which can capture up to 250 frames per second. It is electronically synchronized with the DMD to guarantee that both modulation and exposure are matched. For a 30ms integration, the peak SNR is between 25 and 30dB. Note that, in the current configuration a portion of the normally incident light onto the DMD is discarded as it is sent in the direction of a third arm, symmetric of the relay arm (as shown in Figure 3.2) and currently containing a light trap. It is possible to add a second optical relay and sensor array down this path to implement a programmable dual-band compressive imager.

Finally, the full-field-of-view of this prototype contains an area larger than 720×720 DMD mirrors which are in turn mapped to roughly 300×300 sensor pixels which we can readout with a 2×2 binning. The video stream from the sensor is collected directly by a laptop computer running a specific control software in charge of keeping track of the high-speed pattern measurement process and performing the image reconstruction in real-time. The pixels in the readout frame are then aggregated digitally to synthesize the output of 90×90 *virtual* detectors, each one mapped to a block of 8×8 micromirrors. Finally, we developed an automatic calibration method to infer this mapping and is detailed in a later section.

3.3 Compressive Measurements Design

To write an analytical model of the measurement process, we employ a simple Kronecker linear measurement model where F denotes the discretized scene matrix and is arranged into E blocks (columns) of dimension N (rows) and taking values in $[0, 1]$. The M projection patterns to be used for the compressive measurements are forming the rows of the $M \times N$

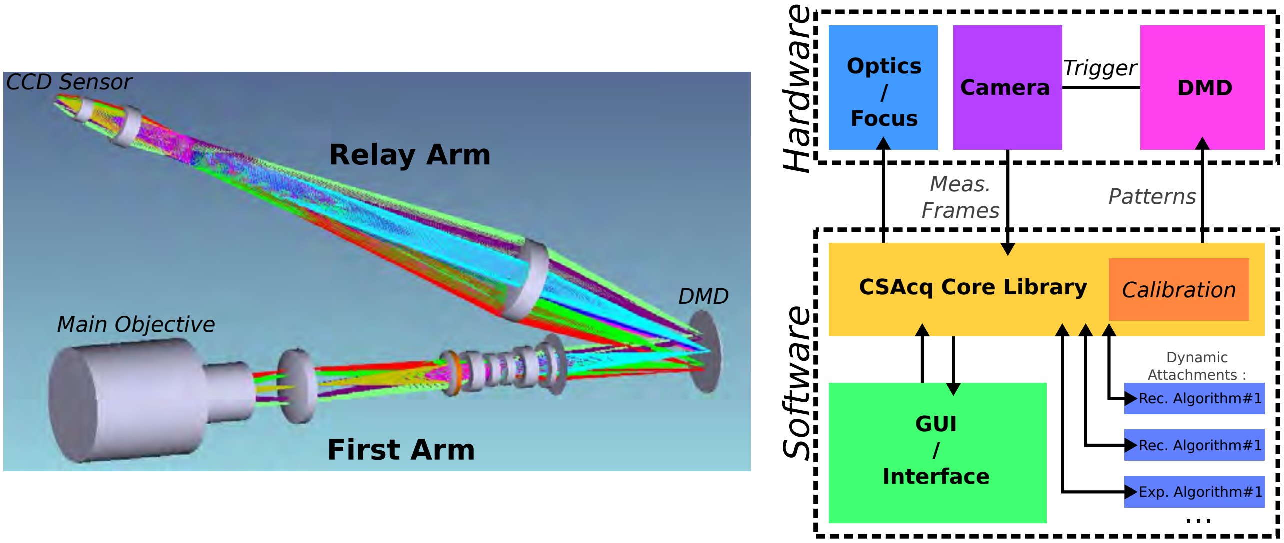


Figure 3.3: **Left** : Custom optical layout showing both the imaging arm and the *relay arm*. **Right** : simplified functional architecture of the prototype and the interface between hardware and software components.

projection matrix P. We define the associated compression ratio $\rho = N/M > 1$ for this projector, for which higher values indicate a higher compression rate. Finally, we note G the $M \times E$ measurements matrix corrupted by additive white Gaussian noise Z of variance σ^2. The linear measurement model can thus be written as :

$$G = PF + Z, \quad \text{with : } \mathrm{vect}(Z) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{EM}), \tag{3.1}$$

where $\mathcal{N}$ denotes the Normal distribution and I_{EM} is the identity matrix in $\mathbb{R}^{EM \times EM}$. Note that the noise variance σ^2 is related to the peak signal-to-noise ratio (SNR) via : $\mathrm{SNR} = -20\log_{10}(\sigma)$. The parallelism of this compressive measurement architecture is thus directly encoded into the matrix product between P and F, as all the columns of F (the blocks in the field of view) are independently projected onto the rows of P (the projection patterns). This model hence differs from the traditional global linear measurement model which is restricted to projecting a single scene vector (*i.e.* $E = 1$). By performing all the scene block projections in parallel, this scene is scalable to large space-bandwidth products : increasing the field of view does not require to add further projections in order to maintain a given compression ratio ρ, only to concatenate more of such blocks. This in turn reduces the hardware constraints, particularly on the frequency of the spatial modulator, as the number of patterns is now independent of the size of the field of view. This whole architecture can be intuitively understood as a concatenation of small and independent single-pixel cameras.

As mentioned earlier, the modulation patterns are the rows of the projection matrix P and are successively displayed by the spatial modulator. These patterns must however first satisfy a *no-gain* constraint as the input light signal cannot be amplified. Secondly, the sum of the maximum exposure (largest modulation coefficients, in absolute value) for each of the patterns must be equal to the exposure of a single frame in a conventional camera in order to establish a fair comparison. Finally, a third constraint arises from the readily available hardware and dictate that the same exposure must be allocated to each and every patterns. All of these can these requirements can be modelled as the following simple constraint on the elements of the sensing matrix $P : \forall i, j \; |P_{i,j}| \leq 1/M$.

For these patterns, a typical heuristic proposes to pick the modulation values from random Gaussian or random signed binary number generators. The restricted isometry property (RIP) [32, 33] then demonstrates that the operator is, with a high probability, preserving or at least limiting the distortion applied to sparse signals. However this approach doesn't consider neither the noisy nature of the measurements nor that the patterns are subject to the previous physical constraints. The result also does not guarantee to reach an optimal compression level. On the contrary, we consider the projections design method presented in by Huang *et al.* in [38], which consists in the optimization of the projection matrix within an information-theoretic framework. The authors use Shannon's mutual information metric J, which quantifies the transfer of information from a statistical model describing the natural scenes to the compressed measurements. They can thus optimize the projection patterns to be used given a fixed measurement noise σ_n and subject to the *photon-count constraint* on the operator. This optimization task can be expressed as :

$$\underset{P}{\arg\max} \;\; J(\mathbf{g};\; \mathbf{f}) \quad \text{s.t.} \quad \forall j, \; \sum_{i=1}^{M} |P_{ij}| \leq 1, \tag{3.2}$$

where the *photon-count constraint* stipulates that the flux, without being amplified, is efficiently used. This is the case of a CS architecture presented in [30]. In this work, we use a series of 8×8 projection patterns optimized within this information-theoretic framework for SNR ranging from 20dB to 45dB SNR, and renormalized to fit the equal-exposure constraint of our hardware platform. These patterns are shown in Figure 3.4.

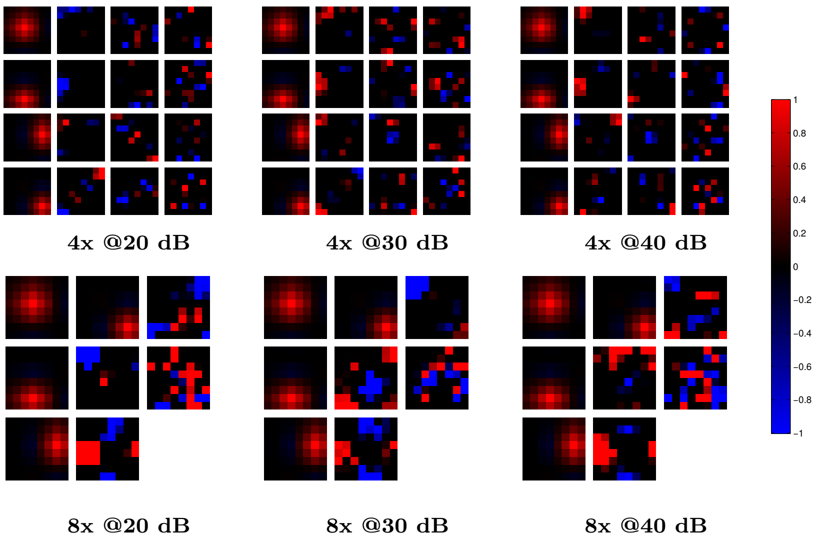


Figure 3.4: 8×8 Information Optimal projection patterns for 4x compression (16 patterns) and 8x compression (8 patterns) for various SNR values. The first four patterns in each (in the left-most column) target low spatial frequencies in the signal while the remaining patterns address high spatial frequencies depending on the SNR budget available. All patterns are normalized to $[-1, 1]$ range for visibility.

As mentioned in the previous section, it is not possible to emulate subtraction in the context of incoherent imaging and as such, we have to split between patterns containing all-positive and all-negative modulation factors. This results in an increase of the corresponding noise variance in Equation 3.1 to at most $2\sigma^2$, depending on the number of signed patterns.

3.4 Scalable Block-wise Reconstruction

Most well-known CS reconstruction algorithm such as $\ell 1$-magic [34], Orthogonal Matching Pursuit [36] or Two-steps Iterative Shrinkage/Thresholding [35], perform a global iterative reconstruction of the measurements acquired with the rank-deficient sensing operator.

Each processes all the data in the field of view at once and sequentially refines the output image under the assumption that the scene signal is indeed sparse. However, this approach results in a time-consuming computation, taking from seconds to minutes to complete, and therefore is not suitable for real-time applications. While these algorithms could be used with the output measurements of the scalable compressive imaging architecture, they would not leverage the parallel nature of the acquisition. Instead, we propose to use a block-wise Bayesian and non-iterative reconstruction algorithm, inspired by the Piecewise Linear Estimator with Maximum A Posteriori (PLE-MAP) presented by Guoshen *et al.* in [37]. This algorithm is especially capable of reconstructing streams at near video frame-rate when running on a common graphics processing units (GPUs).

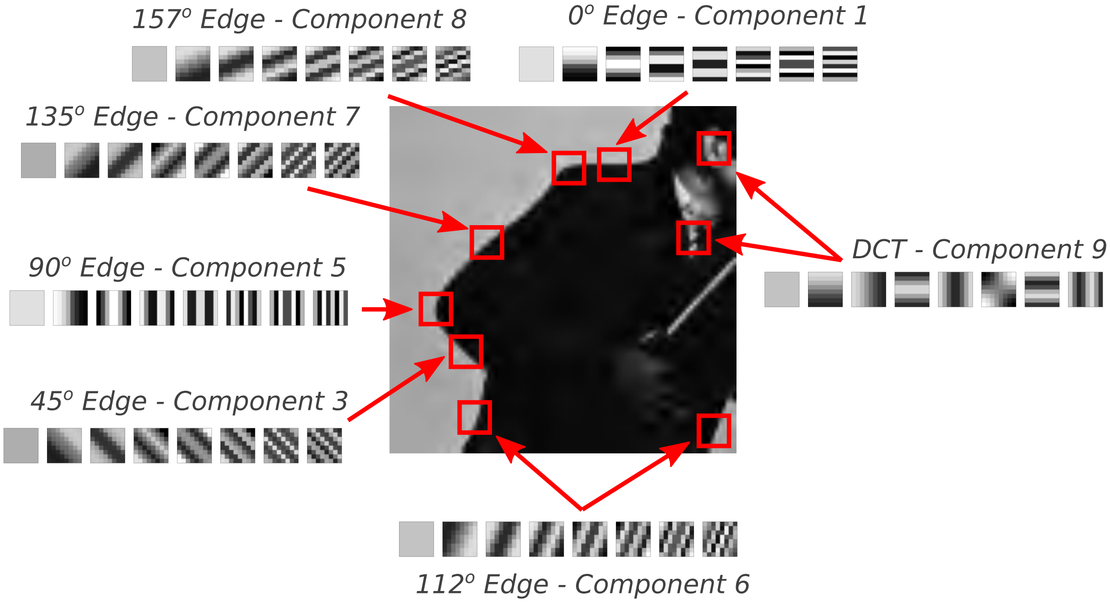


Figure 3.5: How various components of the statistical prior model address specific portions of the image at the block level, and their predominant eigenvectors.

This Bayesian algorithm does not rely on the sparsity property as the previous classical CS algorithms, but on a statistical prior model describing the complex textures of natural images at the scale of a single block. Due to the small block size, those textures depict mostly edges, smooth gradients and a very few complex patterns as illustrated in Figure 3.5. From this observation, we construct a prior distribution $p(\boldsymbol{x})$ of the block content $\boldsymbol{x}$ based on a Gaussian Mixture Model (GMM) for which, each component addresses one edge angle

:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K+1} \omega_k |\Sigma_k|^{-\frac{1}{2}} \exp\big(-\,(\boldsymbol{x}-\boldsymbol{s}_k)^\intercal \Sigma_k^{-1} (\boldsymbol{x}-\boldsymbol{s}_k)\big/\,2\big)\Big/\sqrt{2\pi}^{N}. \tag{3.3}$$

In this expression, each of the $K+1$ multivariate Gaussian components has a weight ω_k, mean vector $\boldsymbol{s}_k$ and covariance matrix Σ_k. We compute these parameters via a procedure similar to that presented in the original PLE-MAP article. First, we start by analyzing the average power spectrum of a dataset of natural images from a standard database, *i.e.* the series of eigenvalues obtained from the average covariance matrix of their blocks. Then, we sample uniformly K angles in the range of $[0; (K-1)\pi/K]$, with K an integer between twice and four times the size of the block (about the size of half to a full perimeter in pixels). For each, we generate a set of synthetic images depicting a single binary edge at that angle and for various translations. We then compute the modes/eigenvectors of this ensemble, *i.e.* the eigenvectors of the covariance matrix to these images blocks, and order them with decreasing eigenvalues. To capture more complex patterns, we also add a single component to the mixture for which the modes/eigenvectors are the vectors of the DCT basis, ordered from low to high spatial frequencies. Each Gaussian component of the mixture is setup with a zero mean ($\boldsymbol{s}_k = \boldsymbol{0}$) and its covariance matrix Σ_k is computed as the product between the eigenvalues of the natural scenes covariance (the power spectrum) and the eigenvectors of a synthetic edge covariance (the modes/eigenvectors). The truncated eigenbasis of a few components are shown in figure 3.6 along with the natural images power spectrum. Finally, all components of the mixture are uniformly weighted : $\forall k,\ \omega_k = 1/(K+1)$. Note that this simplistic prior model is only able to capture structures at small scales, hence it limits the size of the blocks we could use to less than 16×16 pixels approximately.

For the reconstruction algorithm, we adapt the piecewise linear estimator (PLE) from [37] where we replace the maximum a posteriori (MAP) estimator by a minimum mean square error (MMSE) estimator. The PLE-MMSE reconstruction estimator, combined with the use of a GMM prior and the assumption of an additive white Gaussian noise, admits a closed form expression and thus does not require the implementation of a slow iterative refining

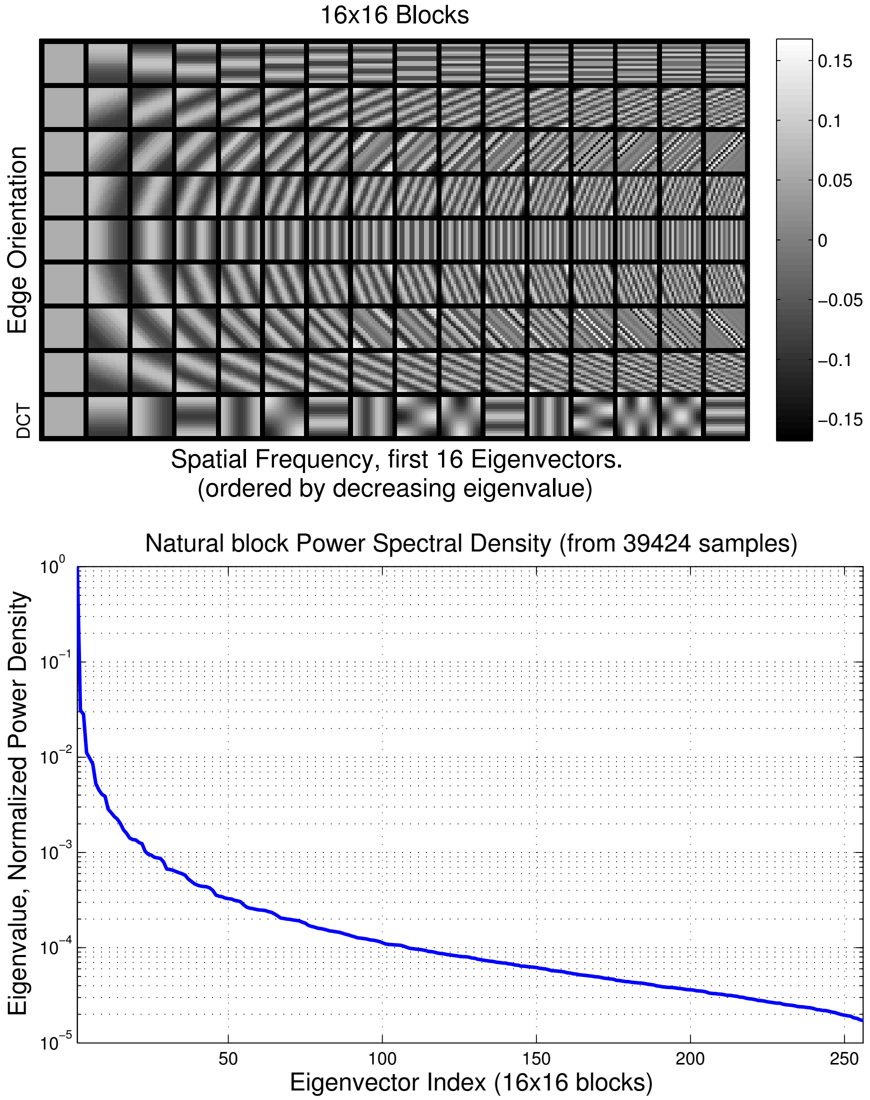


Figure 3.6: **Top** : eigenvectors of some of the components of a coarse prior (few angles) with DCT in the bottom row. They are sorted from left to right in descending order of their spectral power. **Bottom** : the power spectral density (PSD) / eigenvalues shared by all the components of the GMM.

process. This expression can be summarized as follow : first and for each block individually (indexed by e), we reconstruct from the measurements vector $\boldsymbol{g}_e$ one least-square solution $\widehat{\boldsymbol{f}}_{e,k}$ per component of the GMM (indexed by k). This is akin to assuming that the current block does indeed contain the specific type of texture described by a particular mixture component. We write :

$$\widehat{\boldsymbol{f}}_{e,k} = \Sigma_k P^\intercal C_k^{-1} \boldsymbol{g}_e, \qquad \text{with : } C_k = P \Sigma_k P^\intercal + \sigma^2 I_M. \tag{3.4}$$

Also, for each block and each mixture component we compute the following coefficient $w_{e,k}$ indicating how likely it is for the current block to match the texture description given by the mixture component k :

$$w_{e,k} \propto \exp\left(-\boldsymbol{g}_e^\intercal C_k^{-1} \boldsymbol{g}_e / 2\right) |C_k|^{-1/2}. \tag{3.5}$$

Finally for each block, we combine all the reconstructions $\{\widehat{\boldsymbol{f}}_{e,k}\}$ after being weighted by the corresponding $\{w_{e,k}\}$ to obtain the result :

$$\widehat{\boldsymbol{f}}_e = \sum_{k=1}^{K+1} w_{e,k} \widehat{\boldsymbol{f}}_{e,k} \Bigg/ \sum_{k=1}^{K+1} w_{e,k}. \tag{3.6}$$

We note that all of the computationally expensive matrix-products, inverses and determinants of all the matrices $\{C_k\}$ in Equation 3.4 can be calculated in advance as they do not depend on the input signal. This leaves the algorithm only a handful of matrix-vector operations to perform, and such that its complexity is in the order of $\mathcal{O}(EKNM)$: it is linear with respect to the number of blocks E, prior components K, block size N and number of measurements M. From a numerical stability perspective, it is however preferable to compute the natural logarithm of all the weights $l_{e,k} = \ln(w_{e,k})$, and then find and subtract their maximum in order to reduce the overall numerical dynamic range in Equation 3.6.

Nevertheless, as this algorithm reconstructs all the blocks independently from their neighbors, it produces visually disturbing artifacts at the edges of the blocks because the textures can appear discontinuous from one block reconstruction to that of its neighbor. To avoid those, we instead reconstruct all the groups of 2 × 2 adjacent blocks, or *super-blocks*, as a single entities. This is achieved by concatenating both operator and measurements to form a matrix and a vector of size $4M \times 4N$ and $4M$ respectively, as follow :

$$\boldsymbol{g}_{\text{super}} = P_{\text{super}} \boldsymbol{f}_{\text{super}} + \boldsymbol{s}_{\text{super}} \quad \Leftrightarrow \quad \begin{bmatrix} \boldsymbol{g}_1 \\ \vdots \\ \boldsymbol{g}_4 \end{bmatrix} = \begin{bmatrix} P & & \cdots & 0 \\ & P & & \vdots \\ \vdots & & P & \\ 0 & \cdots & & P \end{bmatrix} \times \begin{bmatrix} \boldsymbol{f}_1 \\ \vdots \\ \boldsymbol{f}_4 \end{bmatrix} + \begin{bmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_4 \end{bmatrix} \tag{3.7}$$

With this concatenated model, we have to use a prior generated on 16×16-pixels block following the same method as previously described. We eventually merge the reconstruction results of all these overlapped super-blocks into a smoother final image by averaging the overlapping portions. In Figure 3.7, we show a comparison between the results of the non-overlapped and the overlapped reconstruction methods for the same measurements and same prior mixture. In the overlapped version, the block edges are almost invisible and much finer texture details are revealed. It must be emphasized that this process can be performed without changing the underlying measurement strategy, on 8×8 blocks, and comes only at a the cost of a linear complexity increase for the reconstruction algorithm.

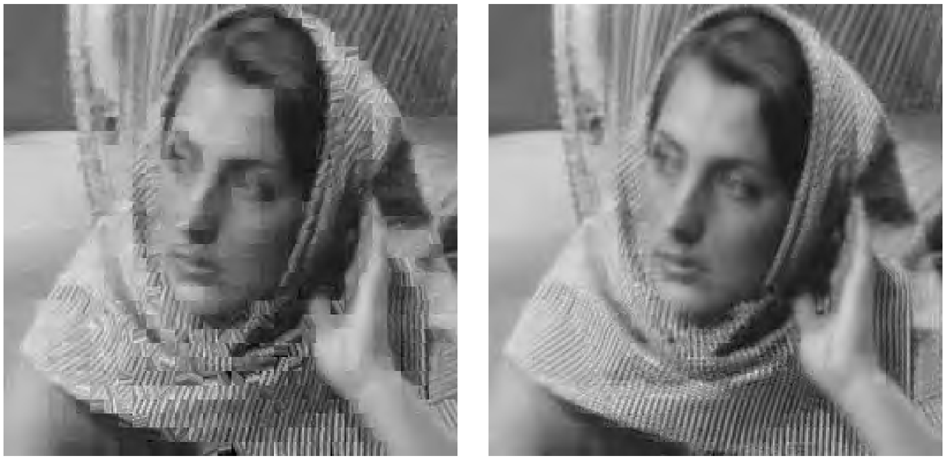


Figure 3.7: Comparison of image reconstruction without (**left**) and with (**right**) overlapping, each using the same set of random binary projection patterns at 4x compression.

Regarding the implementation of the reconstruction algorithm and as noted above, it only requires element-wise and linear algebra operations. It also is natively parallel for all the blocks (or super-blocks) and all the prior components. Thus, it is perfectly suitable for massively parallel computing architectures such as graphic processing units (GPU). Our CUDA® implementation running on a single NVIDIA™ GTX 580 device (in double precision) can reconstruct an image of size 512 × 512 ($\approx$ 0.25 megapixel) in 9 milliseconds, 1024 × 1024 ($\approx$ 1 megapixel) in 35 milliseconds and 4096 × 2048 ($\approx$ 16 megapixels) in 285 milliseconds in agreement with the analytical linear complexity. This task can also easily be split across multiple GPUs, each handling a small region of the field of view in order to obtain a quasi-linear increase in frame-rate.

3.5 System Performance Simulation

As a first evaluation of the algorithm, we simulate the exact analytical expression of the linear measurement model for various inputs. These gray-scale images, with values in $[0, 1]$ are first reordered into a set of 8×8 pixels blocks and projected with a sensing matrix abiding by the previous equal-exposure physical constraint. We then add numerically-generated white Gaussian noise to the projected measurements obtained.

We compare the results generated by random signed binary projections (with values in $\{-1; +1\}$, independently drawn from a Bernoulli process with 50% probability) and Information Optimal projections. We simulate the acquisition and reconstruction from these two operators at 4x and 8x compression ratio (for 16 and 8 patterns, respectively) and while keeping the prior model and other parameters identical. The figures 3.8 through 3.12 show the obtained results for both natural scenes and artificial test targets. Naturally, we observe that for all operators, the quality of the reconstruction is degraded with the increasing compression ratio and decreasing SNR. Nevertheless, the information optimal operators appear capable of retaining most the image features. Moreover, we find that they capture more scene details and generates substantially less artifact in the reconstruction than the random operators. This qualitative observation is supported by the quantitative PSNR values computed between the original object and the reconstructions and presented in Table 3.1. For these examples, we find a gain of up to 5dB at 20dB SNR and 4x compression in favor of the information optimal projector.

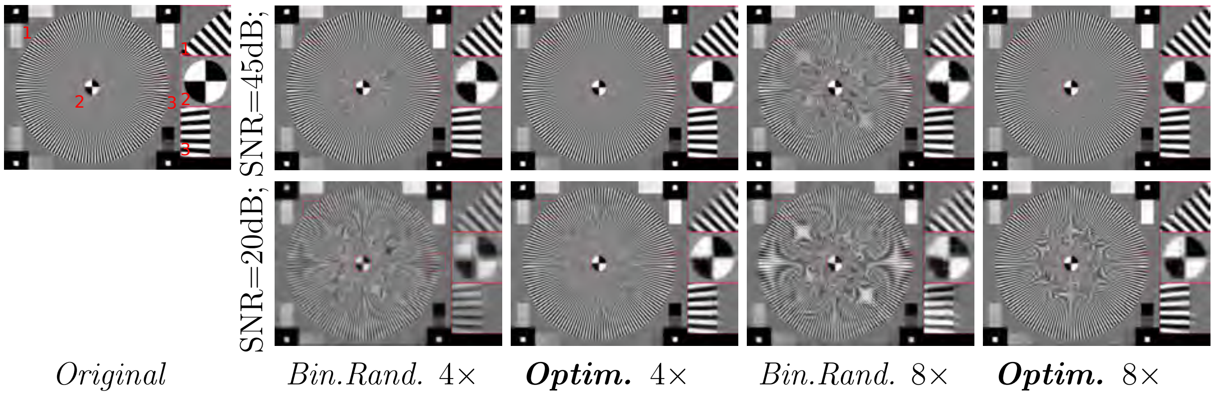


Figure 3.8: **Simulations** for ”*Starchart*” target, comparison of the reconstruction with measurements acquired from binary random $\{-1, 1\}$ and information optimal operators.

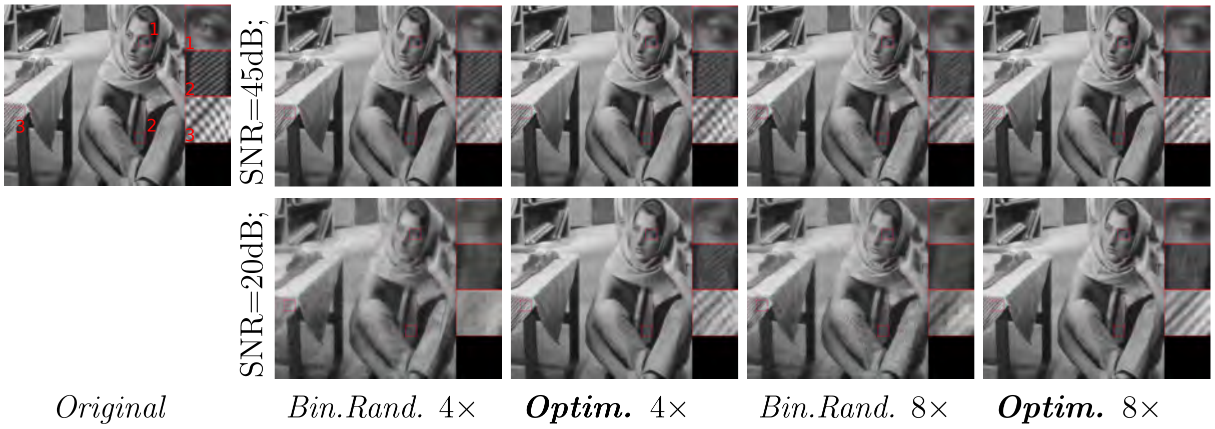


Figure 3.9: **Simulations** for ”*Barbara*” target, comparison of the reconstruction with measurements acquired from binary random $\{-1, 1\}$ and information optimal operators.

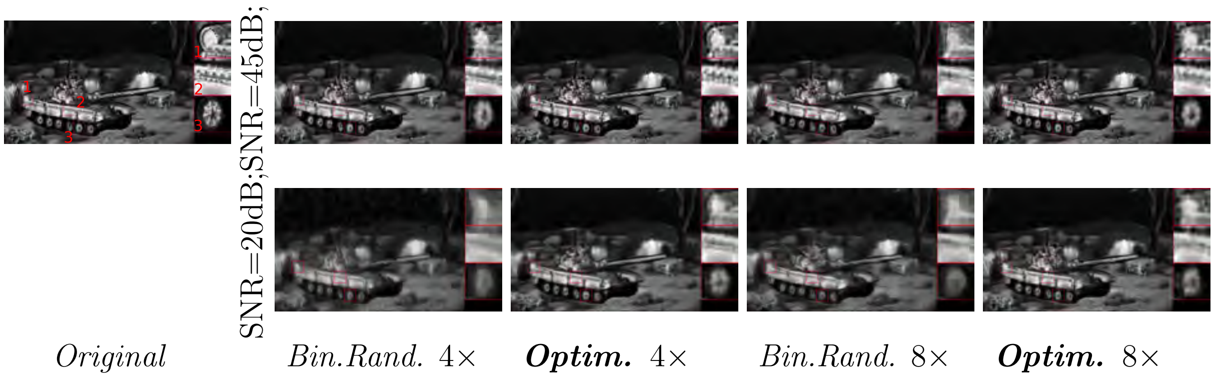


Figure 3.10: **Simulations** for ”*Tank*” target, comparison of the reconstruction with measurements acquired from binary random $\{-1, 1\}$ and information optimal operators.

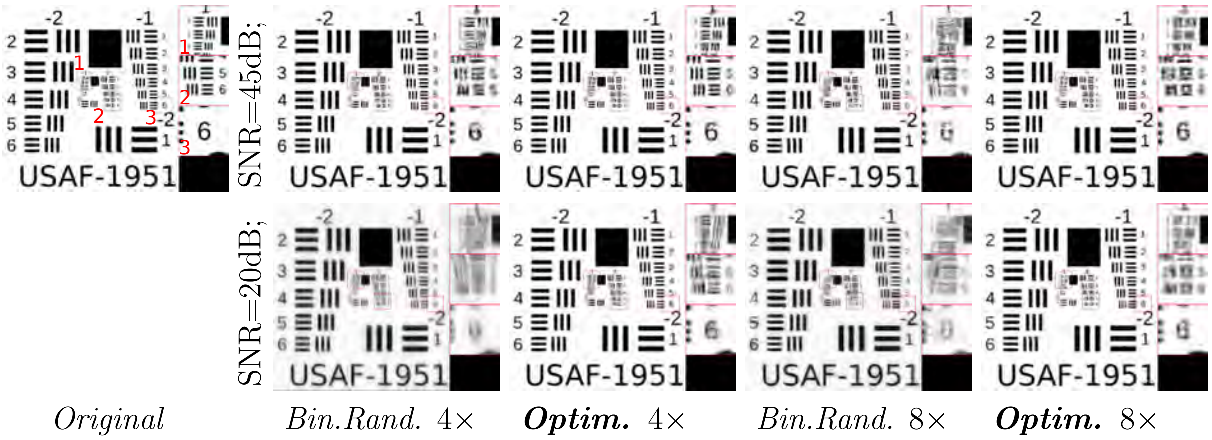


Figure 3.11: **Simulations** for "*USAF bar-chart*" target, comparison of the reconstruction with measurements acquired from binary random $\{-1, 1\}$ and information optimal operators.

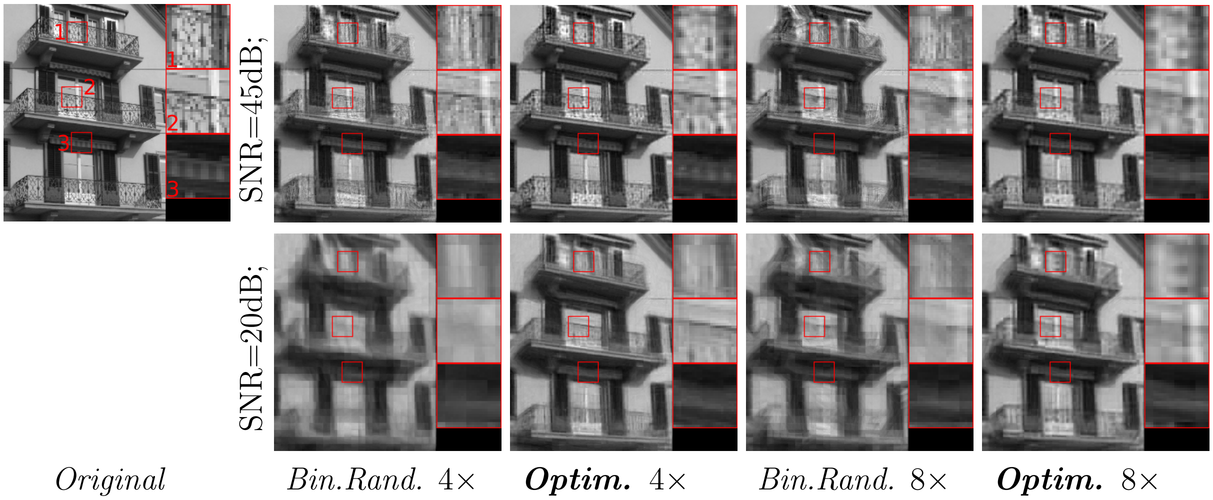


Figure 3.12: **Simulations** for "*Building*" target, comparison of the reconstruction with measurements acquired from binary random $\{-1, 1\}$ and information optimal operators.

PSNR (**dB**)	Random Gaussian				Random Signed Binary				Optimized			
Compression :	4x		8x		4x		8x		4x		8x	
SNR :	*45dB*	*20dB*	*45dB*	*20dB*	*45dB*	*20dB*	*45dB*	*20dB*	*45dB*	*20dB*	*45dB*	*20dB*
USAF-1951	23.31	14.08	20.34	15.27	22.99	15.23	20.30	16.87	**26.55**	**20.74**	**22.37**	**19.64**
barbara	26.89	20.51	**24.29**	21.46	26.68	21.54	24.21	22.25	**27.08**	**24.48**	24.05	**23.25**
Building (training)	23.83	18.08	22.03	19.04	23.59	19.09	22.06	20.09	**25.38**	**22.73**	**23.06**	**21.62**
Starchart	21.21	10.98	16.48	11.33	21.00	11.62	16.82	12.34	**21.92**	**17.41**	**17.88**	**14.67**
Tank	25.06	20.27	23.25	20.98	24.91	21.01	23.25	21.74	**27.14**	**24.05**	**25.16**	**23.56**

Table 3.1: Reconstruction PSNR from **simulation** results at 4x and 8x compressions, 45 and 20dB SNR. The *building* image was used to train the Information-Optimal projections.

3.6 Experimental Implementation

3.6.1 Supervision Software And Calibration Procedure

In addition to the hardware implementation of the architecture described previously, we use a custom software to supervise the calibration, acquisition and reconstruction. This program runs on a single laptop computer and is connected to the DMD, the sensor and an external display panel acting as the scene and facing the system. This computer is in charge of uploading a sequence of global projection patterns to the control board of the spatial modulator. As they are successively displayed on the DMD, a trigger is issued to the camera to match the pattern exposure to the sensor exposure. The recorded image is then transferred back to the computer for processing and the reconstruction results are displayed in a graphical user interface shown in Figure 3.15.

One of the main problem plaguing Compressive Sensing implementations is their acute sensitivity to miscalibration, *i.e.* how a small deviation of the physical measurement process away from the idealized linear model can negatively impact the output quality of a reconstruction algorithm. On these optical architectures we can identify numerous sources of such mismatch. Both **mechanical misalignment** and **optical distortion** prevents the perfect mapping of blocks of modulators to sensors. **Blur** caused by the extent of the relay arm Point Spread Function (PSF) induces signal leakage and cross-talk between neighboring blocks. The physical limitations of the spatial modulators can also pose problem : for instance, the gray-scale projections patterns cannot be reproduced with an infinite precision because of **quantization** of control levels, they also suffer from the **non-linear** and **non-uniform** response of the individual elements (a problem mainly encountered with Liquid Crystal modulators rather than DMDs). Similarly for the sensor itself, the precision of the measurements is reduced by **non-linearities**, **non-uniformities** and **quantization** as well as more complex noise statistics than assumed in our model. Overall, this translates as a significant, but not insurmountable, complexity increase from traditional cameras.

Naturally we can deal with those nonidealities in two ways : either by improving the device and thus the physical measurement model toward the ideal one, or on the contrary generalize the mathematical model to incorporate realistic features of the physical setup. In [44], the authors have presented how it is possible to fold knowledge of the optical characteristics back to the measurement matrix so as to improve the reconstructions. However in the context of this work, we cannot proceed with the same technique as we wish to compare sets of random and optimized projections : for this, we would have to pass the calibration data back to the projections optimization process. Instead, our calibration aims at limiting the alignment and setup defects of the hardware in order to approach the idealized measurement model.

To infer the system parameters, we simply point to a known simple scene and interpret the distortions between the expected and obtained output of the system. In the case of our prototype, we have to learn the association between blocks of micromirrors on the DMD and sensor pixels. However, this mapping cannot be made perfect : the surface of the modulator is indeed seen by the FPA at an angle, with a non-integer scale and is convolved by the PSF of the relay-arm.

In this prototype, we left the modulator to be orthogonal to the optical axis of the imaging arm. At the working F-Number, the geometrical depth of focus is approximately $\Delta \approx 2F_\sharp c$, where c is the diameter of the confusion circle : for the DMD we have about 160μm depth, and only 120μm for the CCD. However, in the configuration we opted for, we can tilt the sensor surface with respect to the optical axis of the relay arm in order to match the image plane conjugated to the modulator surface. While this helps reduce the optical blur, it amplifies the distortion generated by the relay-arm and, thus, the modulators-sensors mapping. Although, as we are forced to virtually perform the reconstruction in the modulation plane, this distortion is not visible in the final reconstructed image.

To setup the system, we first align the optical axis of the first arm and second arm onto the normal of DMD surface with a laser beam. We then position, rotate and focus the sensor while looking at a test pattern displayed on the DMD (typically, a binary checkerboard) and a uniform white screen is facing the first arm. Finally, we focus the first arm objective while observing a textured scene. Thus, the image plane of the first arm, DMD surface and object plane of the second arm, are all coincident, or at least within their respective depth-of-field/focus. Regarding other sources of deviation, we find that the Digital Micromirror device exhibits no significant distortions aside from the mandatory quantization of the gray-scale patterns. We also do not find any significant differences between the acquisition of 8 and 12 bits quantized raw images by the CCD sensor.

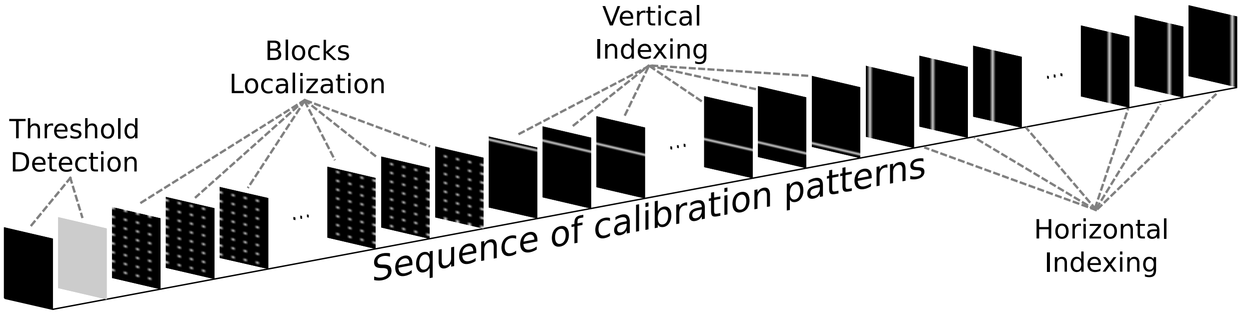


Figure 3.13: Visualization of the binary patterns sequence to be used for the calibration and which cover the entire surface of the spatial modulator. During this procedure we put a uniformly white screen in front of the imaging arm. Hence, as we control the object and the modulation are known, we can infer the mapping through the relay arm and other calibration parameter from the corresponding images recorded by the sensor.

At each observation session, we let the software perform an automated calibration to recognize the mapping data for each of the few thousand blocks composing the field of view. For that, a uniformly white object or screen is once again positioned near the entrance pupil of the first arm. Images are recorded by the sensor while a series of identification patterns are displayed onto the DMD. This sequence is illustrated in Figure 3.13 and composed as follow : first some uniform *on-off* patterns, followed by checkerboards, vertical and horizontal bars; all the same size as that of the block. With this data, the program computes the transition threshold between the two predominant histogram modes of the the uniformly illuminated images. Then, it uses this level to binarize all the remaining images and scans them to detect, localize and index the visible blocks, *i.e.* it infers which pixels of the sensor are mapped to a particular block of modulating elements. Finally, it also calibrates the response of the sensor (bias and linearity) and measures the effective noise standard deviation.

Despite this careful calibration, we find that the optical distortion induced between the DMD plane and the sensor, as well as the extent of the optical Point Spread Function (PSF) of the relay-arm, generates a significant cross-talk between the measurements of these small blocks. The overlap and misalignment of the block images onto the sensor is shown in Figure 3.14. In order to mitigate this effect we split the global pattern patchwork into four, where each block is isolated from its closest active neighbor by at least one fully opaque block. Thus for each projection pattern, the DMD has to display four sub-patterns to observe all of the blocks.

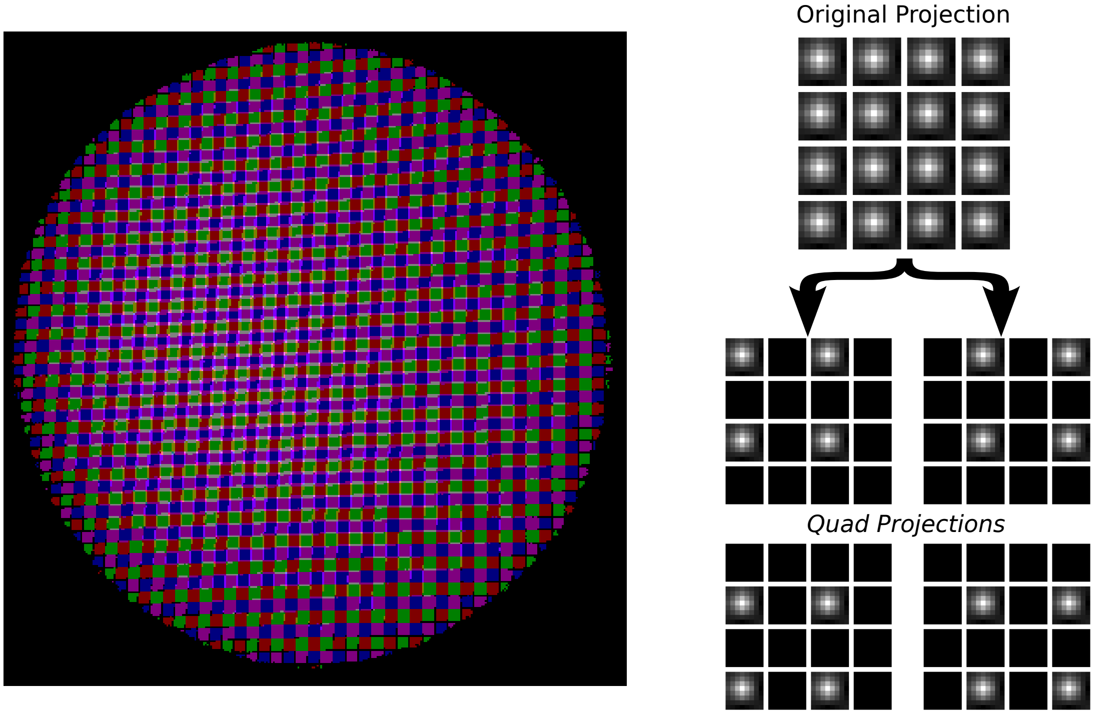
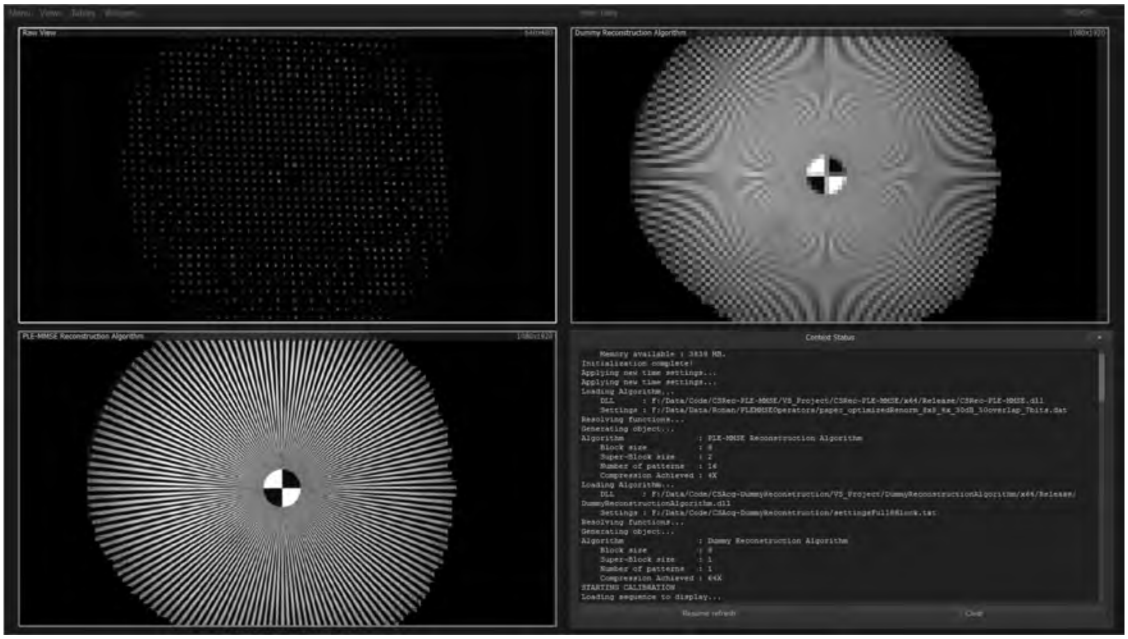


Figure 3.14: **Left** : Visualization of the block segmentation as recorded by the sensor. **Right** : splitting a global pattern (the tiling of a single projection pattern) into 4 sub-patterns to avoid cross-talk from the imperfect mapping onto the focal plane array. The blocks are separated for the visualization only.

After the calibration, the software is able to start performing the measurements. First, the user dynamically pairs it with one or multiple reconstruction algorithms and projection sets. For each, the program builds all the global patterns and upload them to the DMD. During the acquisition, it extracts the measurements by scanning the raw sensor images it receives with the previous calibration data : it averages the pixel values mapped to a particular block into a single scalar. This way, it progressively fills the appropriate measurement matrix G of Equation 3.1. Once all of the patterns were displayed, the sequence is looped back to the first. Thus the system can deliver a continuous video stream with each of the previous matrix being used as a circular buffer, *i.e.* receiving steady updates. In the meantime, the running reconstruction processes are asynchronous with the acquisition of the compressed measurements : this allows the software to deliver an output image with a smaller period than that of filling completely the matrix G.


Menu Views Tables Widgets
Raw View
640x480
Dummy Reconstruction Algorithm
1080x1920
PLE-MMSE Reconstruction Algorithm
1080x1920
Context Status
Resume refresh
Clear

```
    Memory available : 3838 MB.
Initialization complete!
Applying new time settings...
Applying new time settings...
Loading Algorithm...
    DLL      : F:/Data/Code/CSRec-PLE-MMSE/VS_Project/CSRec-PLE-MMSE/x64/Release/CSRec-PLE-MMSE.dll
    Settings : F:/Data/Data/Ronan/PLEMMSEOperators/paper_optimizedRenorm_8x8_4x_30dB_50overlap_7bits.dat
Resolving functions...
Generating object...
Algorithm              : PLE-MMSE Reconstruction Algorithm
    Block size         : 8
    Super-Block size   : 2
    Number of patterns  : 16
    Compression Achieved : 4X
Loading Algorithm...
    DLL      : F:/Data/Code/CSAcq-DummyReconstruction/VS_Project/DummyReconstructionAlgorithm/x64/Release/
DummyReconstructionAlgorithm.dll
    Settings : F:/Data/Code/CSAcq-DummyReconstruction/settingsFull8Block.txt
Resolving functions...
Generating object...
Algorithm              : Dummy Reconstruction Algorithm
    Block size         : 8
    Super-Block size   : 1
    Number of patterns  : 1
    Compression Achieved : 64X
STARTING CALIBRATION
Loading sequence to display...
```

Figure 3.15: Acquisition and Reconstruction software interface, showing in real-time the raw compressed measurements stream (**top-left**), the image captured by the conventional system with the low-resolution sensor (**top-right**) and the reconstruction output for the PLE-MMSE algorithm (**bottom-left**). Reconstruction and other exploitation algorithms can be added dynamically to the program.

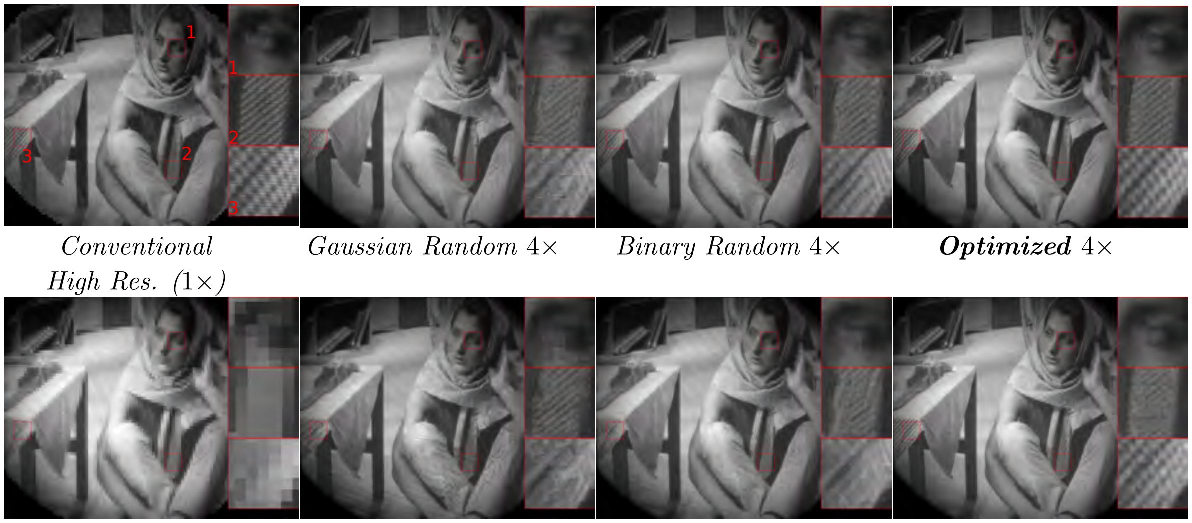
3.6.2 Experimental Results

For the experimental tests, the compressive camera is recording the display showing test images previously used for simulation. These gray-scale images are precorrected by the inverse gamma function $x \mapsto x^{\frac{1}{2.2}}$ to compensate for the dynamic range compression applied by the monitor and so that the output luminance is linear. Note that we also have to make sure that the pulse-width modulation of the backlight (for brightness control) does not produce temporal aliasing artifacts with the fast acquisition. Finally, we ensure to use the same calibration in order to nullify differences between imaging sessions and to allow for a fair comparison of the results. The final images are all reconstructed via the PLE-MMSE block algorithm using the same 64 angle priors, DCT prior, and 2×2 super-block overlapping technique.

In addition to the compressed measurements we also emulate two configurations of the conventional camera which represents either the highest or lowest performance bounds. The first imitates a high-resolution sensor in place of the spatial modulator : it is implemented by measuring the pixel (or canonical) basis. This operator consists of 64 projections, each with only a single element (pixel) of the 8×8 block in the *on*-state and thus it does not achieve compression (1x). The second imitates the low-resolution sensor used to acquire the block projections although put in place of the spatial modulator. This is implemented by only measuring the block DC component, where all the elements are in the *on*-state, and is equivalent to an effective 64x compression from the previous high-resolution. For both, we perform a simple linear back-projection for the reconstruction.

As for the simulation, we compare the results obtained for random signed binary and Information Optimal projections. We also add properly normalized random Gaussian projections (with values drawn independently from a Gaussian process). All these operators were first normalized with the equal-exposure constraint such that each of their patterns is contained in $[-1; 1]^{8\times 8}$ and quantized to 7 bits in order to be displayed by the DMD (the least significant bit of the transmission value is removed). The measurements obtained exhibit a peak-SNR between 25 and 30dB and the reconstructed images are approximately 700×600 pixels in size.

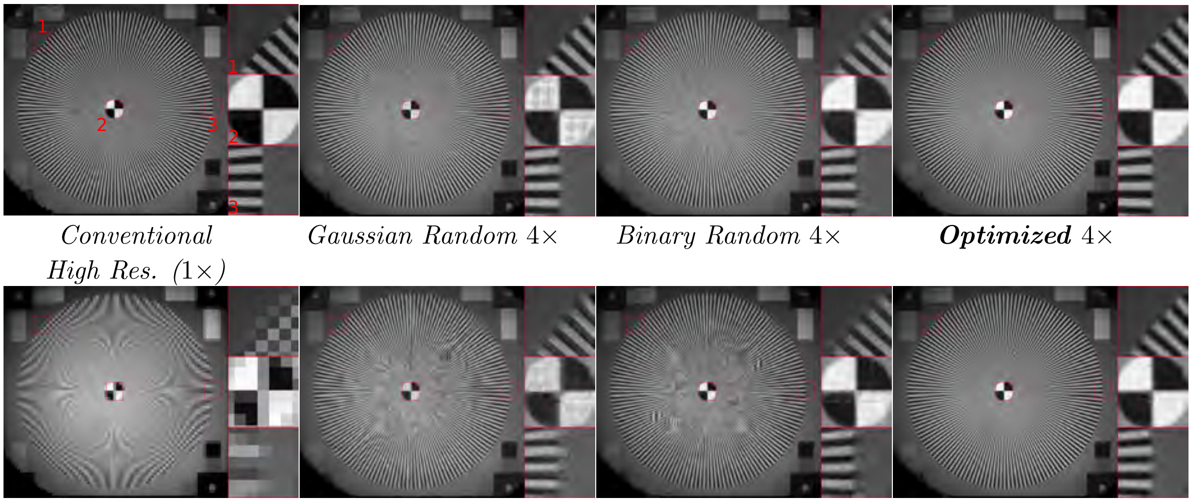
The resulting reconstructions for all the operators are shown in figures 3.16 through 3.20 and tend to corroborate the observations reported for the simulation results. From the resolution test targets, one can verify that the information optimal projections deliver images of higher effective resolution than both Gaussian and signed binary random projections while acquiring only a fraction of the measurements of a conventional camera. This is particularly visible at a compression rate of 8x where characters and smaller bars of the USAF test chart are still distinguishable. For natural images, we observe that the small-scale features, such as the clothes textures in Figure 3.16, are also better preserved by the information optimal projections.



Conventional High Res. (1×) *Gaussian Random 4×* *Binary Random 4×* ***Optimized** 4×*

Conventional Low Res. (64×) *Gaussian Random 8×* *Binary Random 8×* ***Optimized** 8×*

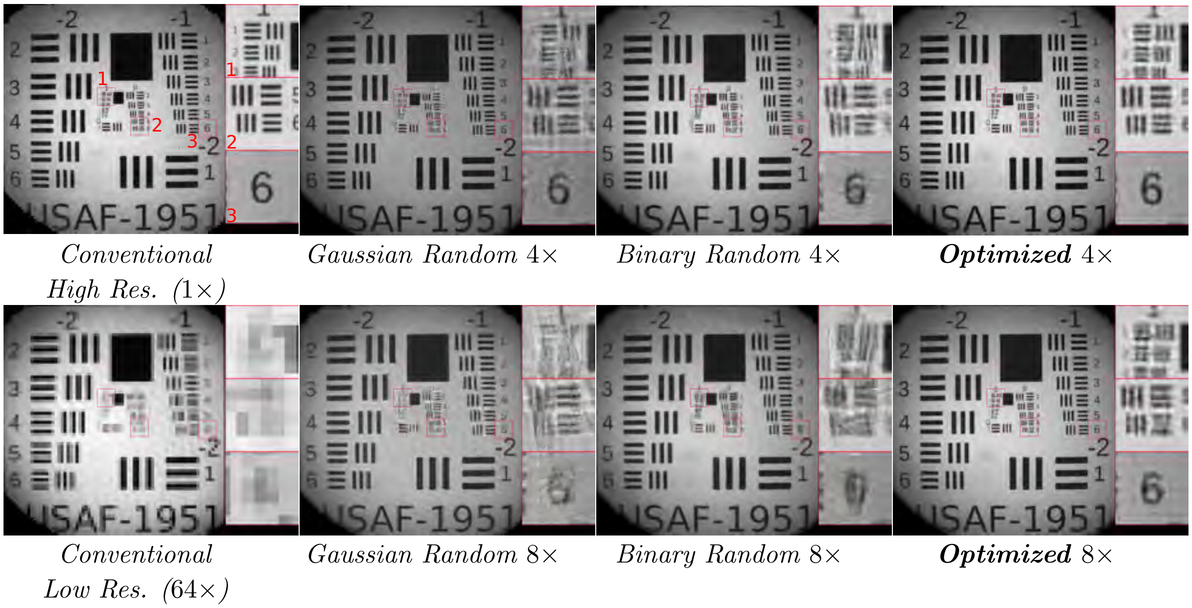
Figure 3.16: **Experimental results** : Barbara image.



Conventional High Res. (1×) *Gaussian Random 4×* *Binary Random 4×* ***Optimized** 4×*

Conventional Low Res. (64×) *Gaussian Random 8×* *Binary Random 8×* ***Optimized** 8×*

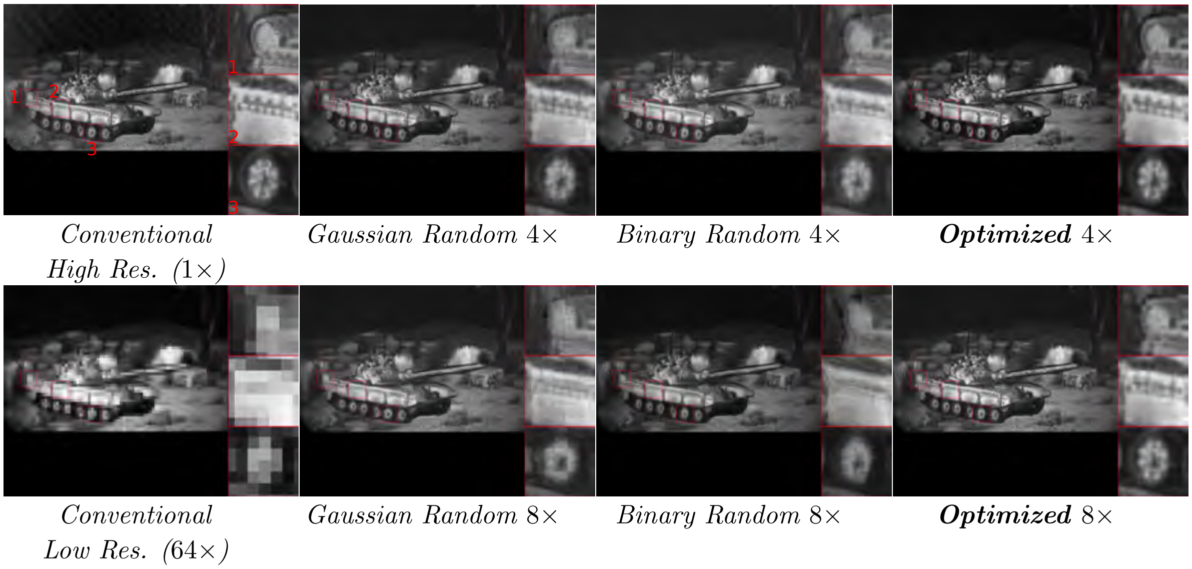
Figure 3.17: **Experimental results** : Star-chart.



Conventional High Res. (1×) *Gaussian Random* 4× *Binary Random* 4× ***Optimized*** 4×

Conventional Low Res. (64×) *Gaussian Random* 8× *Binary Random* 8× ***Optimized*** 8×

Figure 3.18: **Experimental results** : USAF-1951 test target.



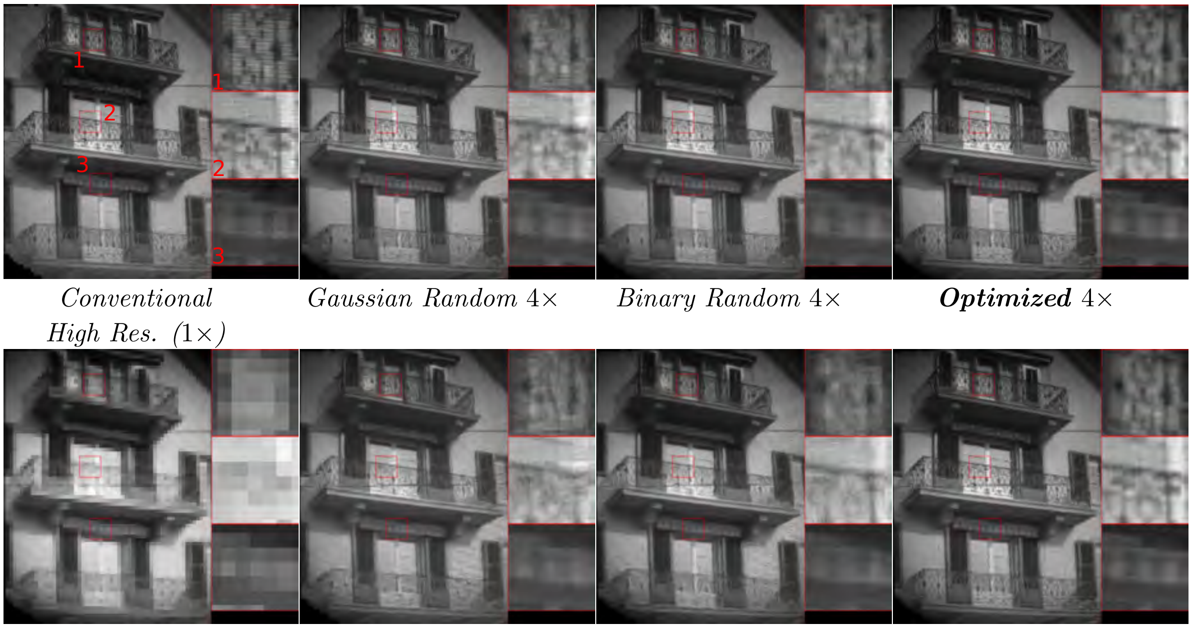
Conventional High Res. (1×) *Gaussian Random* 4× *Binary Random* 4× ***Optimized*** 4×

Conventional Low Res. (64×) *Gaussian Random* 8× *Binary Random* 8× ***Optimized*** 8×

Figure 3.19: **Experimental results** : Armored vehicle.

3.7 Conclusion

We have described the implementation of a parallel compressive imaging architecture which is inherently scalable to high space-bandwidth products, and a similarly parallel non-iterative Bayesian reconstruction algorithm based on a piece-wise statistical model of natural



Conventional High Res. (1×) *Gaussian Random 4×* *Binary Random 4×* ***Optimized** 4×*

Conventional Low Res. (64×) *Gaussian Random 8×* *Binary Random 8×* ***Optimized** 8×*

Figure 3.20: **Experimental results** : Building, image initially used for the optimization of the projections.

textures. This compressive imager is capable of reconstructing megapixel-size images at near video frame-rate when other CS architectures, such as the SPC, are constrained to use slow iterative reconstruction algorithms not suitable for real-time operations. While this platform, as well as the other CS architectures, are more sensitive to miscalibration than conventional imagers, they present significant advantages over the latter : as they leverage the sparsity of natural images, they can operate below Nyquist's sampling limit and thus can employ low-resolution sensors. They also perform fewer measurements than a traditional imaging architecture and thus require a smaller readout bandwidth and benefit from a multiplexing gain advantage over the direct measurements. Finally, while our device prototype is working in the visible domain, it is easily adaptable to SWIR/MWIR by swapping the FPAs. It is also possible to dispose of the second arm, which is unused in our current configuration, to build a programmable dual-band observation system. In thermal infrared however, the background radiation is lowering the modulation contrast which affects the resulting image quality. Hence, a CS camera working in this particular domain faces additional engineering challenges, such as described in [69].

As an application of this scalable compressive imager, we have demonstrated that the information-optimal projections deliver superior image quality when compared to both random Gaussian and random signed binary operators in simulations as well as in real-world conditions with our prototype implementation. This performance gain is particularly remarkable at high compression (8x) or low SNR (25dB) where mean square error PSNR can reach a gain of up to 5dB in favor of the optimized measurements.

While we have used our prototype as a testbed for the straightforward imaging configuration, it is fully programmable and automated. Therefore, it can be employed to benchmark any other projection operators, reconstruction and exploitation algorithms [45, 46]. Among these more advanced capabilities, we can explore adaptive acquisition strategies, *i.e.* designing the future projection patterns from the knowledge of the previously acquired measurements, which can help dramatically increasing the compression ratio. We can also implement a foveated acquisition strategy by changing the projections used only for some specific portions of the field of view. This way, we can imagine focusing on a few regions of interest and allocating minimal resources elsewhere. Finally, in the context of machine vision, we envision that this system can be used to not only acquire measurements and reconstruct an image of the field of view, but also to acquire direct information about targets in the scene.

Chapter 4

Scalable Compressive Target Detection and Classification Sensor

The programmable implementation of the scalable compressive camera described previously is not limited to only imaging tasks. In this chapter, we discuss how the same architecture can be used to perform automatic target detection and recognition without reconstructing the intermediate scene image. To improve the overall performance of this system, we develop an optimization framework of the projection patterns based upon the Cauchy-Schwarz mutual information. We especially prove that this metric provides an upper-bound to the probability of misclassification.

4.1 Introduction

Recent advances in electro-optic/infrared (EO/IR) based automatic target detection/recognition (ATD/ATR) have witnessed a significant increase in the field of view and spatial resolution, especially for airborne platforms [50, 51]. This has lead to increasing image data that in turn puts severe strains on the design and implementation of conventional imaging optics, focal plane sensors and subsequent image exploitation algorithms. This has significant implications for resource costs associated with the focal plane array (FPA), especially in infrared spectral regime, in terms of manufacturing complexity, readout bandwidth, power consumption for real-time operation. In parallel, the emergence of compressive imaging architectures has demonstrated that non-traditional imaging systems offer an alternate imaging paradigm with favorable system resources scaling compared to traditional imaging architectures [30]. Instead of sampling the scene irradiance field densely, *i.e.* at Nyquist sampling, compressive imaging system exploit the compressibility of natural scenes to acquire fewer measurements leading to reduced FPA density, lower readout power and smaller data stream. One of the well-known implementation of a compressive imaging architecture is the Single Pixel Camera (SPC) [31], which spatially modulates the scene irradiance field before integrating it onto a single photodetector. By performing successive linear projections, from a library of spatial modulation patterns (e.g. random binary patterns), the system can form an image of the entire scene computationally using nonlinear image reconstruction algorithms. As this architecture is built around a single photo-detector it represents lower cost alternative to the more expensive high-resolution FPA employed in a conventional imaging system, especially in short and mid-wave infrared spectral bands. However, the SPC architecture is subject to its own scalability limitations. For example, the rate of spatial modulator increases quadratically with the FOV and spatial resolution which limits the total number of measurements that can be collected within a fixed exposure time. Furthermore, the complexity of the reconstruction algorithm increases exponentially with the image resolution which limits real-time operation. To overcome this measurement and computational scaling issue, we have developed a parallel scalable compressive imaging architecture capable of delivering megapixel-scale images in near real-time. This scalable architecture exploits a low resolution sensor array, where each pixel integrates a small region (*i.e.* a block) of spatially modulated scene irradiance to implement a single pixel camera. Thus the sensor array effectively implements an array of SPC operating in parallel, each imaging a region of the FOV. The compressive measurement from the FPA are then processed by a non-iterative (single shot) reconstruction algorithm with a complexity linear with respect to the resolution. For image formation applications, we have demonstrated that this compressive imaging system can reconstruct the same field of view than a traditional imager but with a lower resolution FPA that has up to 64× fewer pixels and operating at a compression ratio between 4× to 8×.

It is important to note that the aforementioned compressive imaging systems and their acquisition strategies are optimized for the task of image formation. While it is indeed possible to use the reconstructed image as an input to an image exploitation algorithm for the task of ATD/ATR, it does not represent optimal approach to maximizing the task performance. This is due to the fact that for specific tasks, such as target detection and recognition, the underlying information of interest in a scene is quantifiably smaller than that for the image formation task[57, 59, 60]. Thus by carefully designing the linear projections, incorporating appropriate task and scene priors, it is possible to selectively preserve the relevant task information with much fewer compressive measurements, and thus achieve higher compression ratio relative to traditional compressive image formation followed target detection and recognition. This task-specific compressive measurement design approach has been explored by several researchers in the literature: Dunlop-Gray *et al.* have implemented a simple classifier for a compressed hyper-spectral imager [67]; Baheti *et al.* [58] and Huang *et al.* [68] have designed information optimal adaptive compressive measurements to a face recognition task against a large face database. In [66] Li *et al.* demonstrated that a SPC paired with Principal Component Analysis (PCA) Secants projections, further optimized via semidefinite programming, can recognize simplistic targets over a uniform background at a higher rate than performing inference on image estimated from compressive random projection and also direct inference on random projection measurement.

In this work, we focus on the development and implementation of a scalable compressive approach to target detection and classification where the target of interest may be embedded in diverse natural terrains. The inherent variabilities of the target appearance coupled with the complexity of the background scene suggests that sophisticated statistical models are required to faithfully model such targets. This presents a two fold challenge in terms of the design and optimization of the projection patterns for compressive measurements, as well as formulating and implementing the target detection/recognition algorithm that operates on the compressed data directly, without the need for intermediate image reconstruction. Furthermore, the compressive measurement system design must also scale to wide field of views if the proposed system is to be of any practical utility. To satisfy these scaling requirements, we adopt a compressive measurement architecture that is based on a block-wise decomposition of the scene, similar to our scalable compressive imaging architecture described in chapter 3. This approach allows us to consider groups of blocks that specify a minimum region of interest and can measured compressively (in parallel) for direct target detection and recognition.

This work is organized as follows: in section 4.2 we begin with a description of the concept of operation of the proposed compressive imager followed by a comprehensive description of the underlying statistical model of the targets and the background scene in terms of the Gaussian Mixture Model (GMM). Then we describe the mathematical model of the scalable compressive imaging system in section 4.3, which acquires the compressed measurements of the scene. Those are directly used by the classification algorithm, described thereafter, to perform the target detection and recognition task. In section 4.4 we will describe an information-theoretic framework for compressive measurement design along with the corresponding design metric and how the measurements will be optimized based on the scene statistical model and other system parameters such as the measurement signal to noise ratio (SNR) and measurement compression ratio. In section 4.5 we present a simulation analysis of the proposed compressive measurement system performance for the target detection and recognition task, followed by an experimental validation of the simulation predictions. Finally we present our conclusions in section 4.6 along with a brief discussion of future directions of work.

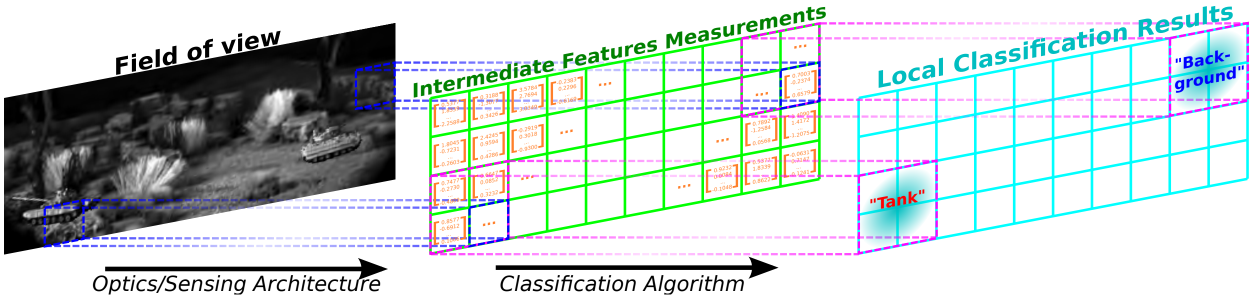

Field of view
Intermediate Features Measurements
Local Classification Results
"Back-ground"
"Tank"
Optics/Sensing Architecture
Classification Algorithm

Figure 4.1: Architecture abstraction : the field of view is divided into contiguous square block-tiles. The device optics and sensing components measure in parallel a small features vector for each. Finally, every group of 2 × 2 blocks is then analyzed by a classification algorithm which determines a decision label for the corresponding location.

4.2 Scalable Scene Model

4.2.1 System Operation Concept

The task of detecting the presence and classifying the type of vehicles visible in a large field of view has already been thoroughly investigated across multiple direct acquisition modalities, such as visible and infrared wavelengths or Synthetic Aperture Radar (SAR), due to its critical importance for surveillance applications. However, the signal acquisition and processing have been considered as disjoint parts of the system imaging pipeline. They have therefore been designed separately on the model of traditional imaging architectures for the first and standard image processing concepts for the second. While this approach is seemingly flexible, it comes at the cost of having to capture, readout and process images of large dimensionality in WFOV applications. For example, in the case of an autonomous aircraft surveilling a wide area, the implementation of a direct acquisition scheme for ATD/R requires to use a large FPA sensor to record images with a sufficiently fine ground resolution in order to distinguish the targets of interest. The large images acquired must then be compressed and transmitted for remote exploitation. Even with specialized hardware, all of the on-board operations consume a significant amount of power to acquire, process and compress the data. Once received and decompressed, the image is first passed to a fast detection algorithm which marks the portions of the scene containing a target. This first stage is usually implemented as a correlation filter sweeping the whole FOV, as it is the case for the matched filters [49]. Those recorded hits are then further examined by a second algorithm to classify the target type. For high-resolution images, this data processing and exploitation requires a significant amount of computational power to be performed in real-time.

While Mahalanobis *et al.* have show that this two-stage classifier can be adapted to the output compressive imager [69], we are here interested in bypassing the intermediate scene reconstruction step which is a resource-intensive and time-consuming operation. The compressive sensing architecture we use can particularly be designed and optimized from end-to-end and exclusively for the particular task at hand. In Figure 4.1, we show the abstract acquisition concept : the field of view is seen as a grid of contiguous blocks by the joint optical and detection architecture, which acquires a feature vector in parallel for each block. These measurements are then directly passed to a detection and classification algorithm, without performing a superfluous image reconstruction. The program independently analyzes the feature vectors for each patch of 2×2 blocks, or chip, and issue a class label simultaneously determining the presence and the type of the target. As stated previously, it is important to note that the classification performance of this system is the only metric of interest and that the projection patterns measuring the feature vectors should be optimized for this particular purpose only.

For the current demonstration, it is assumed that the scene can contain vehicles among two known positive classes : either Tank or Armored Projection Vehicle (APV). The natural environment surrounding these is composed of dense clutter and is considered as a third negative class. To simplify the context, we assume that the targets are always visible from open-grounds and never occluded by any natural element. The target vehicles are also small in comparison to the field of view but slightly larger than the size of a chip which we fix to 64×64 pixels (*i.e.* 32×32 pixels blocks). Finally, each chip can be labeled as *background* (target absent), *tank* or APV and we assume that these three cases are equiprobable. In the remainder of this section, we describe the construction of a statistical model describing

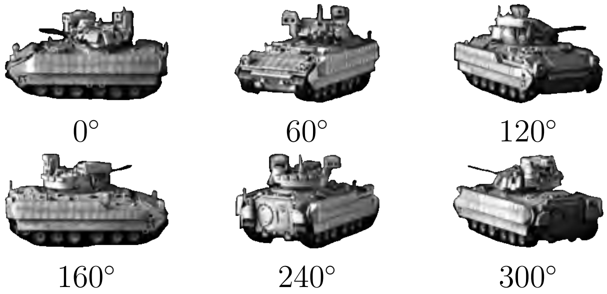


Figure 4.2: High-Resolution samples for the class #1 : model APV.

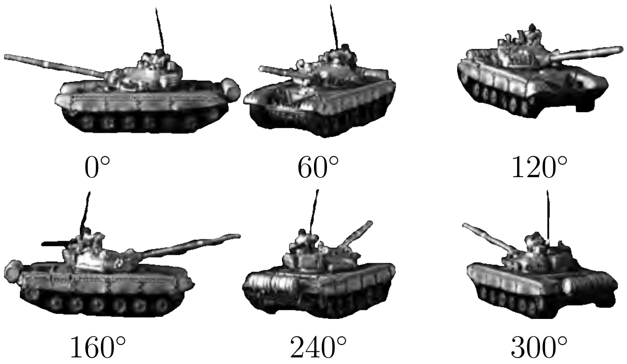


Figure 4.3: High-Resolution samples for the class #2 : model T72.

these scenes in order to (a) serve as ancillary data for the classifier and (b) optimize the projection patterns used via an Information-Theoretic framework.

4.2.2 Target And Background Model

To generate a dataset of the scene chips, we first acquire high-resolution images of miniature vehicles in various orientations as shown in Figures 4.2 and 4.3 and of natural scenes with different configurations as shown in Figure 4.4. They are used to generate synthetic chip samples by transforming and overlaying a vehicle masked-image on top of a terrain image. To produce realistic scenes, the transformations can consist in any combination of : translation (in 2D, as a uniform grid of 16×16 positions in the chip which includes the vehicle being occluded outside of the chip), in-plane rotation (±15°, in three discrete increments), out-of-plane rotation (full 360°, in discrete 20° increments) and scaling (30% range

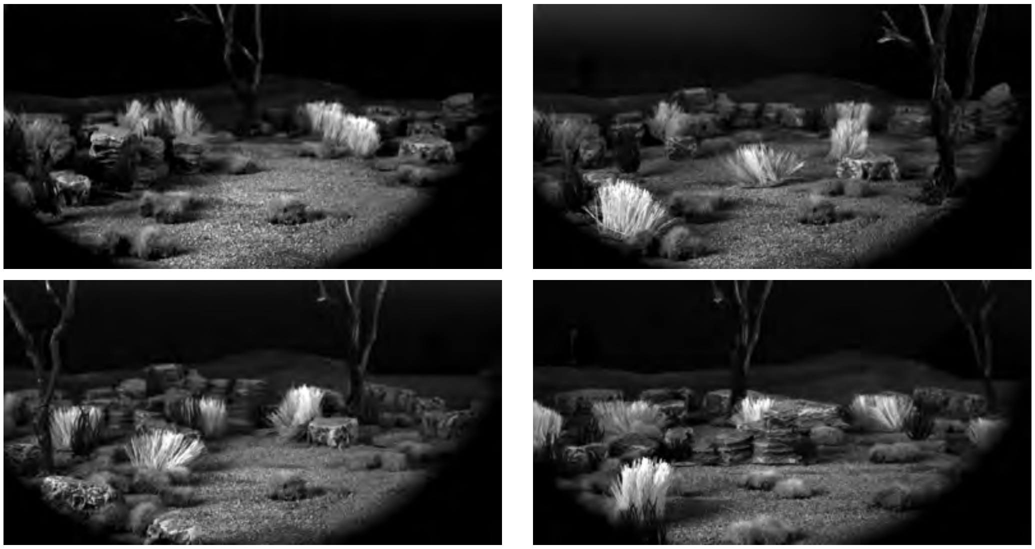


Figure 4.4: Background scene samples.

in two increments, to emulate the varying distance range). Finally, the target samples are appropriately blurred in order to match their Power Spectral Density (PSD) to that of the backgrounds to avoid the presence unnatural and overly-salient artifacts in the composite image.

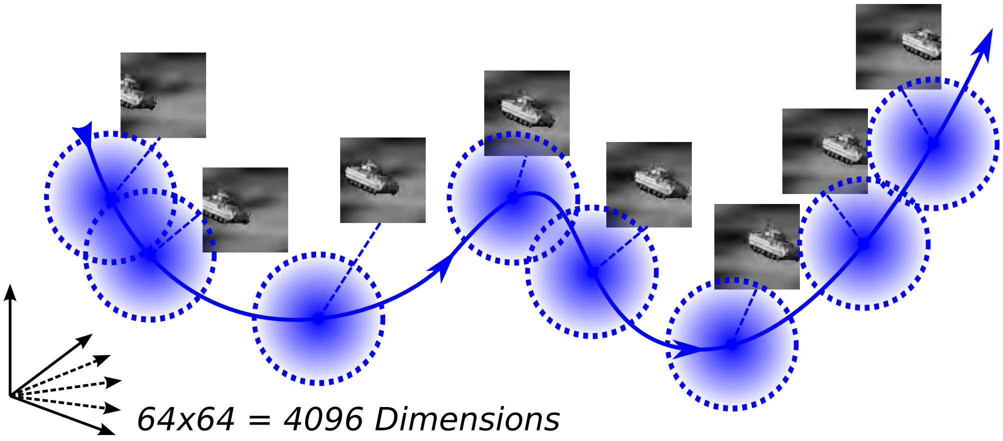


Figure 4.5: Simplified visualization of the manifold sampling for a single degree of freedom (here, the translation of the target vehicle). The Gaussian kernels will partly fill up the space left between the samples.

In a noiseless measurement environment, a class dataset over a particular terrain can be pictured as a low-dimensional manifold embedded in a 4096-dimensional space which is obtained by continuously varying the parameters of the previous transforms for both the target and the terrain. It is however impossible to find the exact analytical expression of this complex surface. Instead, we construct an approximated statistical model of a target class by sampling points from this manifold and using them as the component means of a mixture of Gaussians. As illustrated by Figure 4.5, the gap existing between neighboring components is then filled by the extent of the Gaussian distributions which is governed by their covariance matrices. In the current context, we setup all the mixture components with the same weight and the same spherical covariance matrix of scalar variance σ^2. Therefore, we can write the likelihood distribution for the class variable C taking the label-value i of a chip sample $\boldsymbol{f}$ as :

$$p(\boldsymbol{f}|C=i) = \frac{1}{O_i}\sum_{k=1}^{O_i}\mathcal{N}\left(\boldsymbol{f}|\boldsymbol{t}_{i,k}, \sigma^2 I_{EN}\right), \tag{4.1}$$

where O_i is the total number of components in the mixture, $\boldsymbol{t}_{i,k}$ is the k^{th} sample from the manifold belonging to the class i. Here, we define the constants E as the number of blocks in a chip and N as the dimension of a block, *e.g.* in the the current context we have $E = 2^2 = 4$ and $N = 32^2 = 1024$. We note that the vectors $\boldsymbol{f}$ and $\{\boldsymbol{t}_{i,k}\}$ are in a $EN = 4096$ dimensional space, more precisely we constrain the chip images to the normalized hypercube volume $[0,1]^{4096}$ (*i.e.* the dynamic range of the grayscale values is between zero and one). Finally, I_{EN} denotes the identity matrix in a space of EN dimensions. For the mixture corresponding to background scenes, we can employ a similar modeling strategy and uniformly sample its manifold. However, in order to reduce the number of components in the mixture for this negative class, we only select 4096 centroids generated by the K-Means clustering algorithm over a large number of background realizations. It is important to remark that the total number of components O_i of any target mixture and the variance σ^2 are related : if the manifold is coarsely sampled then this variance must be large enough to fill the wide gaps between the samples. In that case, the smaller structures of this surface are lost to the approximation. To avoid this, we can construct large mixtures of Gaussians with more than one hundred million components for each target class by uniformly sampling the parameter ranges listed previously. With this fine sampling the standard deviation σ

for our targets is about 0.15. However, those distributions require a large amount of storage space. To address this situation, we first remark that these mixtures enumerate the pairs of every possible target attitude and background centroid and we can therefore dissociate them by storing the masked target samples separately from the background samples. We then assemble the component means when needed and only have to store mixtures containing about twenty-six thousand components for the largest. Figure 4.6 shows a few example of component means found in the statistical likelihood distributions which we are going to use for the optimization of the projections and as ancillary reference data for the classification algorithm. For testing purpose, we also create a dataset of realistic scenes by randomly sampling the target and background manifolds. Note that the test samples thus produced are not found directly among the means of the class likelihood mixtures.

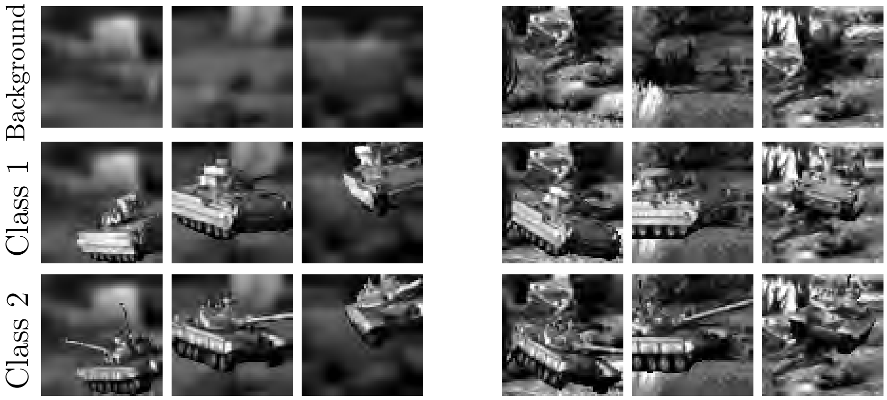


Figure 4.6: 64 × 64 Chips examples, as 2 × 2 blocks of 32 × 32 pixels. The attitude (set of transform parameters) of the targets is kept constant in each column. *Left* : Training samples (for pattern optimization and classification reference). *Right* : Testing samples.

4.3 Compressive System Design

4.3.1 Scalable System Model

Following this description of the chips statistical model, the linear measurement model must also adapt to the block wise nature of the acquisition. This is the case of our compressive sensing architecture which can be represented as a concatenation of many smaller SPCs operating in parallel, and therefore it can be scaled to high space-bandwidth products (*i.e.* resolution × field of view) by adding more of these independent units.

This block wise measurement model can be expressed with basic linear algebra operations : we first define $\boldsymbol{f}$ a $\mathbb{R}^{EN}$ vector representing the discretized image of a single chip. The constants E and N are defined similarly to the previous section : E denotes the number of blocks constituting a chip ($= 4$) and N the dimension of each block ($= 1024$). The blocks are each sequentially projected against a unique series of M modulation patterns which are included inside the rows of a $M \times N$ projection matrix P. The compression ratio ρ associated to this projection is equal to the blocks dimensionality divided by the total number of patterns : *i.e.* $\rho = N/M$, and in order to achieve significant compression the number of patterns M must remain much lower than the dimension N of the canonical space. Finally, we note $\boldsymbol{g}$ the vector in $\mathbb{R}^{EM}$ representing output measurements obtained for a chip and we assume that they are corrupted by an additive white Gaussian noise $\boldsymbol{x}$ of variance σ_n^2 : *i.e.* $\boldsymbol{z} \sim \mathcal{N}(0|\sigma_n^2 I_{EM})$. These random fluctuations are affecting the acquisition for uncooled sensors and are in part due to the thermal agitation of the charge carriers. The peak Signal-to-Noise Ratio (SNR) of this process is directly obtained from the variance as : $\mathrm{SNR} = -20 \log_{10}(\sigma_n)$. With these notations, the linear measurement model can be written as :

$$\boldsymbol{g} = (I_E \otimes P)\boldsymbol{f} + \boldsymbol{z}, \tag{4.2}$$

where $\otimes$ is the Kronecker product and I_E is the identity matrix in $\mathbb{R}^{E \times E}$: the projection operation onto the rows of the matrix P is thus repeated on each block. One can also transform the previous expression to avoid computations over many empty matrix-blocks by folding the terms as follow : we note F the $E \times N$ matrix obtained by appropriately rearranging the vector $\boldsymbol{f}$ so that each block of the chip occupies a column; similarly, we note G and Z the $E \times M$ matrices associated respectively to the vectors $\boldsymbol{g}$ and $\boldsymbol{z}$. With these, Equation 4.2 can be modified to recall the usual linear measurement model :

$$G = PF + Z, \tag{4.3}$$

where P acts on all columns of F in parallel. While the choice of the projection patterns is explored in section 4.4, it is important to highlight here the constraints arising from the physical implementation of this architecture. More precisely as the modulation of the incoherent light intensity is passive, the projection matrix P cannot apply gain to the input signal. Consequently, all of the values for the matrix P, *i.e.* the patterns modulation coefficients, must remain smaller than 1 in absolute magnitude. Moreover, the spatial modulator device to be used in the prototype will display the patterns successively, each with the same temporal exposure. Thus, in order to establish a fair comparison between operators having different number of patterns and against a traditional camera using the total available exposure, we must further constrain the coefficients down to the range $[-1/M, 1/M]$. Therefore, one can remark that increasing the number of patterns leads to a reduction of the amount of energy allocated to each of them. From this observation we expect to find a compromise over the number of projections M for a fixed block size N, balancing between a too severe compression (small M/large ρ) and an inefficient use of the input flux (large M/small modulation amplitude). In addition, this optimal allocation also depends on sensor noise variance σ_n^2 : in the case of a high σ_n^2, it is preferable to use a smaller number of projections having each binary transmissions (their modulation coefficients are either ± 1) as they will help maximize the incident flux on the detectors and increase the effective signal to noise ratio; on the opposite, for lower σ_n^2 it becomes possible to employ many more projections with more complex gray-scale modulations.

With the measurement model of Equation 4.2, we can now project the statistical likelihoods introduced previously, from the canonical space to the compressed space by using

Bayes' Theorem :

$$p(\boldsymbol{g}|C=i) = \int_{\mathbb{R}^{E\times N}} p_Z(\boldsymbol{g}|\boldsymbol{f})p(\boldsymbol{f}|C=i)d^{EN}\boldsymbol{f} \tag{4.4}$$

$$= \frac{1}{O_i}\sum_{k=1}^{O_i}\mathcal{N}(\boldsymbol{g}|\boldsymbol{s}_{i,k} = (I_E \otimes P)\boldsymbol{t}_{i,k},\ S_{i,k} = I_E \otimes (\sigma^2 PP^\intercal + \sigma_n^2 I_M)). \tag{4.5}$$

This transformation can be decomposed into three operations : (*1*) the means $t_{i,k}$ of each of the components are projected into the $\mathbb{R}^{EM}$ space, (*2*) each of the (spherical) covariances is *squeezed* into the image space of P and (*3*) the overall projected mixture distribution is blurred by the kernel of the additive noise. Especially, this last operation will erase some of the smallest details from the compressed mixture.

Similarly to the measurement model, one can derive a folded, or rearranged, version of Equation 4.5 from the properties of Kronecker product in order to avoid storing and computing large and empty matrix-blocks :

$$p(\boldsymbol{G}|C=i) = \frac{|\Sigma|^{-\frac{E}{2}}}{O_i\sqrt{2\pi}^{EM}}\sum_{k=1}^{O_i}\exp\left(-\frac{1}{2}\operatorname{Tr}\left[(G-Q_{i,k})^\intercal P^\intercal \Sigma^{-1}P(G-Q_{i,k})\right]\right), \tag{4.6}$$

where $Q_{i,k}$ is the $\mathbb{R}^{E\times N}$ block-folded matrix corresponding to the vector $\boldsymbol{s}_{i,k}$, Tr denotes the trace operator and $\Sigma = \sigma^2 PP^\intercal + \sigma_n^2 I_M$ is the squeezed covariance. Note that this last term is common to all the components and is only dependent on fixed parameters of the device : the projection operator P, the model variance σ^2 and the noise variance σ_n^2. Therefore, the inverse and determinant of this positive semi-definite matrix can be computed once and for all.

4.3.2 Target Detection And Classification Algorithm

After the acquisition, the measurements are passed as inputs to a detection and classification algorithm. This often implemented as a two-stages process over the high-resolution image for a conventional system : the first perform a fast detection analysis and the second classifies the region of interest marked by the first. Applying the same method on the measurements captured by the compressive architecture requires to first perform a full reconstruction of the scene from the measurements. However, this operation comes with a significant computational cost and will not provide more information about the target than available from the measurements. Therefore, we can bypass this step and perform the simultaneous detection and classification directly with the acquired features vectors. For this, we write the analytical Maximum-A-Posteriori (MAP) classifier $\widehat{c}_{\mathrm{MAP}}$ acting on the chip measurements vector $\boldsymbol{g}$ (*e.g.* the concatenated 2×2 blocks features measurements) :

$$\widehat{c}_{\mathrm{MAP}}(\boldsymbol{g}) = \arg\max_{c} \left\{ p(c|\boldsymbol{g}) \propto p(\boldsymbol{g}|c)p(c) \right\}. \tag{4.7}$$

This classifier only uses the compressed class likelihoods $p(\boldsymbol{g}|c)$ given by Equation 4.5 and the class priors $p(c) = 1/N_c$ where N_c is the number of classes. This estimator is also of interest as it is guaranteed to minimize the probability of error (or misclassification) P_e which is formally defined for any classifier $\hat{c}$ as :

$$\mathrm{P_e} = p(\widehat{c}(\boldsymbol{g}) \neq c_{\text{ground-truth}}) \leq 1 - \frac{1}{N_c}. \tag{4.8}$$

The inequality given in Equation 4.8 corresponds to the worst possible error-rate and is attained for a random classifier uniformly picking among the N_c possible labels. We add that the definition of the probability of error $\mathrm{P_e}$ can also be expanded into Equation B.10 of the Appendix and relates to an integral over the minority domains of the class likelihoods. While the probability of error is the main metric for the system performance, it is also useful to consider the confusion matrix D of the system in order to understand the interferences between the class labels. The elements of this matrix are defined as the probability for the classifier to issue the label i for an element of the class j, *i.e.* : $D_{i,j} = p(\widehat{c} = i | c_{\text{ground-truth}} = j)$. Note that the sum of the elements along any column is equal to unity : *i.e.* $\forall j,\ \sum_i D_{i,j} = 1$. Finally, the error-rate is related to the coefficients outside of the main-diagonal (*i.e.* the probabilities of misclassification) via : $\mathrm{P_e} = \sum_{i \neq j} p(C = j) D_{i,j}$, where $p(C = j) = 1/N_c$ for the equiprobable classes in the current context.

The computational complexity of the classification algorithm expressed by Equation 4.8 is linear in the total number of components across the mixtures distributions, *i.e.* $\mathcal{O}(\sum_i^{N_c} O_i)$. Nevertheless, the class likelihood distributions presented previously contains more than two hundred million components and therefore, this algorithm would be too slow to operate in real-time. However, we find that the targets only occupy a small portion of the chip (typically, less than 25%) and consequently, the background contributes to most of the signal energy. Hence, the previous classifier can be approximated by dividing it into two stages : the first finds the most likely background component from the chip measurements, and the second is a MAP classifier which only considers components of the targets mixtures tied to the previous background. This method reduces the complexity of the algorithm to $\mathcal{O}(O_b + \sum_{i \neq b}^{N_c} O_i/O_b) = \mathcal{O}(\sum_{i \neq b}^{N_c} O_i/O_b)$, where O_b is the number of components in the background mixture and is equal to 4096 in the current context.

4.4 Measurements Design And Optimization

As established previously, the selection of the projection patterns making up the rows of the projection operator P is critical to the overall system performance in terms of the probability of error defined in Equation 4.8. Several common strategies can be considered here. Random Gaussian projections are certainly a recurrent choice in Compressive Sensing as they benefit from theoretical results, such as the restricted isometry property (RIP) [32, 33] or the Johnson-Lindenstrauss lemma [47, 48], guaranteeing that the linear compression applied will limit the distortion of the signal to less than some fixed threshold. However, it also prevents to adequately compress the signal by failing to exploit the prior information and focusing instead on image reconstruction.

Instead, it is desirable to use the classes prior information to improve the projection design specifically for the task at hand. For instance, the PCA-Secants algorithm [65, 66] constructs the most prevalent projection patterns between samples of two classes. To do so, it first computes the covariance from all the unit-length vectors directed between any two elements from different classes; then it diagonalizes the obtained matrix and selects the eigenvectors corresponding to the largest eigenvalues similarly to the principal component analysis. More intuitively, these secant vectors capture the distribution of the normals to the boundary between the classes as illustrated in Figure 4.7. Therefore, they form good candidates for the projection operator. However the PCA-Secants algorithm misses a few important properties. Firstly, it does not offer an explicit relationship to the previous probability of misclassification and secondly, it is not taking the measurement noise (or acquisition SNR) into account. Finally, it does not support physical constraints of the projection patterns, such as the equal-exposure requirement described previously.

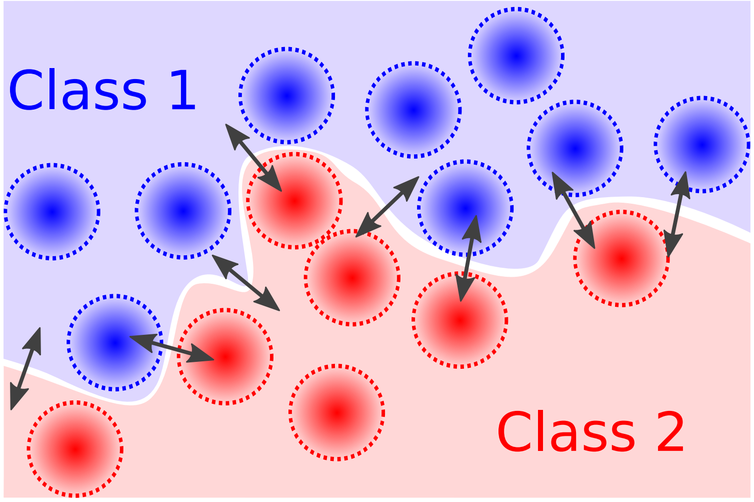


Figure 4.7: Example of local normals to the boundary between two classes and the components of the respective mixtures.

Naturally, the pattern selection process should be cast as an optimization problem which minimizes the error rate, subject to the appropriate projection constraints. While this $\mathrm{P_e}$ metric is ideal, it does not typically admit a closed form expression and its numerical computation requires to finely sample the class likelihoods distribution for an accurate integration, which is computationally intractable. This challenge is further supplanted by the lack of a smooth gradient expression, forcing the optimization to rely on less efficient routines to search the vast operator space of dimension $M \times N$.

Instead, it is common to rely on information metrics that provide bounds to the probability of error. One of the most well-known information-theoretic metric is Shannon's mutual information [52], which imposes on $\mathrm{P_e}$ both a lower bound via Fano's inequality [53] and an upper bound via Kovalevskij's inequality [54]. However, this metric can only be written in closed form for elementary distributions and not for the statistical mixtures we envision for the current problem. Furthermore, its gradient also does not have a closed form expression for the sophisticated statistical model of the scene. While not directly useful for our optimization process, it is interesting to notice that this particular information score can be perceived as the limit case $\alpha \to 1$ of Renyi's Generalized mutual information [55]. The case $\alpha = 2$ leads to the Quadratic mutual information (QMI) which measures the $\mathcal{L}_2$ distance between the joint distributions of class labels and measurements, and the product of their marginals [62]. In this case, the metric and its gradient can be developed analytically for mixtures of Gaussians but only offer a lower-bound to the error rate [61] which has limited utility in our optimization framework.

We thus need an information-Theoretic metric that can be related to an upper bound of the probability of misclassification. In the following section we describe one such metric : the Cauchy-Schwarz mutual information. While this metric has been been studied and employed to optimize and characterize other measurements problem with Mixtures of Gaussians [63], Poisson processes [64] and Mixtures of Poisson distributions [70], we present here several properties justifying regarding the particular classification task at hand.

4.4.1 CSMI Metric

The Cauchy-Schwarz mutual information (CSMI) is directly related to the Cauchy-Schwarz Divergence $\mathrm{D_{CS}}$ which measures the angle between two probability distributions which are square integrable. In the current context, let a discrete class variable C take values in the labels set $\mathcal{C}$ and vector measurements variable $\boldsymbol{G}$ in $\mathbb{R}^{EM}$, their CSMI is defined as :

$$\begin{aligned}\mathcal{I}_{\mathrm{CS}}(\boldsymbol{G}, C) &= \mathrm{D}_{\mathrm{CS}}(p_{\boldsymbol{G},C}(\boldsymbol{g}, c),\ p_C(c)p_{\boldsymbol{G}}(\boldsymbol{g})) \\ &= -\ln\left(\frac{\sum_{c\in\mathcal{C}}\int_{\mathbb{R}^{EM}} p_{\boldsymbol{G},C}(\boldsymbol{g}, c)p_{\boldsymbol{G}}(\boldsymbol{g})p_C(c)\ d^M\boldsymbol{g}}{\sqrt{\sum_{c\in\mathcal{C}}\int_{\mathbb{R}^N} p_{\boldsymbol{G},C}(\boldsymbol{g}, c)^2\ d^M\boldsymbol{g}\sum_{c\in\mathcal{C}} p_C(c)^2\int_{\mathbb{R}^N} p_{\boldsymbol{G}}(\boldsymbol{g})^2\ d^{EM}\boldsymbol{g}}}\right), \qquad (4.9)\end{aligned}$$

where the numerator can be interpreted as the dot product between the joint distribution $p_{\boldsymbol{G},C}(\boldsymbol{g}, c)$ and product of marginals $p_C(c)p_{\boldsymbol{G}}(\boldsymbol{g})$, while the denominator relates to the product of their respective $\mathcal{L}_2$ norms. While the angular separation is explicit in the previous expression, it is possible to further simplify it with the class likelihood distributions. We can here assume that all the classes are equiprobable (*i.e.* $p(c) = 1/N_c$) and we define the system overlap class matrix $V \in \mathbb{R}^{+N_c\times N_c}$, for which each element is written as an overlap integral of two likelihoods :

$$V_{i,j} = \int_{\mathbb{R}^{EM}} p(\boldsymbol{g}|C = i)p(\boldsymbol{g}|C = j)\ d^{EM}\boldsymbol{g}. \qquad (4.10)$$

From there, we can draw a parallel between this matrix, made of positive coefficients, and the confusion matrix previously defined which is related to the probability of error. Both of them share a similar structure : the goal of the optimization is to increase the on-diagonal values which are linked to the probability of correct classification events, while decreasing all the off-diagonal values which represents the likelihood of misclassification (or overlap between different classes). Hence, both of the matrices are diagonal in the case of a perfect classifier; and both are proportional to a matrix of ones in the case of a totally confused classifier (when $\mathrm{P_e} = 1/N_c$). Note however that their normalizations differ : while the confusion matrix has the sum of the values along each of its column equal to unity, the overlap matrix is subject to both the triangular and Cauchy-Schwarz inequalities, *i.e.* respectively : $\forall i, j,\ 2V_{i,j} \leq V_{i,i} + V_{j,j}$ and $V_{i,j}^2 \leq V_{i,i}V_{j,j}$. To further illustrate the underlying function of the CSMI metric, we can highlight that the on-diagonals elements of the overlap matrix are linked to the *within-class* tightness, *i.e.* they represent a measure of how tightly packed are the mixture components of a single class likelihood. The off-diagonal elements, on the contrary, are measuring the distance between mixture components belonging to distinct classes.

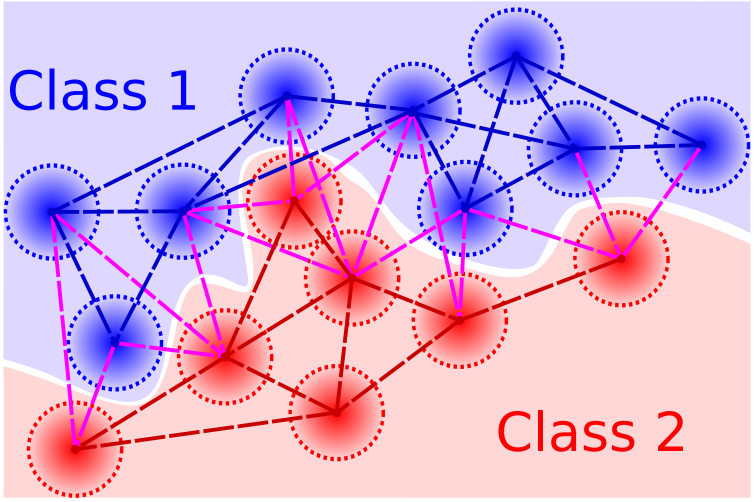


Figure 4.8: Example of component pairs in a two-classes problem : the CSMI metric aims at reducing the distance between components belonging to the same class (either blue or red links) while increasing the overall separation of the two classes (*i.e.* increasing the length of the magenta links).

With this construct, the CSMI expression given by Equation 4.9 can now be simplified for equiprobable classes to :

$$\mathcal{I}_{\mathrm{CS}}(\boldsymbol{G}, C) = \frac{1}{2}\ln(N_c) - \frac{1}{2}\ln\left(1 + \frac{\sum_{i=1}^{N_c}\sum_{j\neq i}^{N_c} V_{i,j}}{\sum_{i=1}^{N_c} V_{i,i}}\right). \tag{4.11}$$

Note that $\mathcal{I}_{\mathrm{CS}}$ value ranges from 0, if all the $V_{i,j}$ are equal, to $\ln(N_c)/2$ if $V_{i\neq j} = 0$. Within this simplified CSMI expression, the coefficients $V_{i,j}$ can be developed into analytical expressions when the class likelihoods are mixtures of Gaussians, by observing that the integral of the product of two multivariate Gaussians can be written in closed form. By using Equation 4.4 of the projected distributions, we have :

$$V_{i,j} = \frac{1}{O_i O_j}\sum_{k=1}^{O_i}\sum_{l=1}^{O_j}\mathcal{N}((I_E \otimes P)(T_{i,k} - T_{j,l})|\, \mathbf{0}, 2I_E \otimes (\sigma^2 P P^{\intercal} + \sigma_n^2 I_M)), \tag{4.12}$$

which can be interpreted as a sum of the pair-wise overlap between all Gaussians components of two mixtures. The development of this expression as well as the gradient of CSMI with respect to the projection operator P are detailed in section B.2 of the Appendix. Thus, we can observe that increasing the magnitude of the main diagonal elements ($V_{i,i}$) or reducing the magnitude of the off-diagonal elements ($V_{i \neq j}$) in the overlap matrix is equivalent to increasing $\mathcal{I}_{\mathrm{CS}}$. Consequently, increasing the CSMI metric is equivalent to *pulling together* components belonging to the same class (*i.e.* increasing all $V_{i,i}$) while *pushing away* components of any other classes (*i.e.* increasing all $V_{i \neq j}$).

Finally, the CSMI metric also provides an upper-bound $\mathrm{U_{Pe}}$ on the probability of error for the classification task, when the likelihoods are modeled by mixtures of Gaussians. Specifically, in the current context where all components share the same weights and covariance matrix, we can write the following inequality :

$$\mathrm{P_e} \leq \mathrm{U_{Pe}} = \frac{\max\{O_i\}}{2} \sqrt{(N_c - 1)\left[\exp(\ln(N_c) - 2\,\mathcal{I}_{\mathrm{CS}}) - 1\right]}. \tag{4.13}$$

A generalized expression of this upper-bound for arbitrary components weights and covariances is also given in section B.3 of the Appendix. While the value of $\mathrm{U_{Pe}}$ is likely not indicative of the true error rate, it is interesting to see that it is equal to zero when CSMI is equal to its maximum value. Hence, a small increase of the metric can provide a valuable decrease in $\mathrm{P_e}$ especially near the vertical asymptote at $\mathcal{I}_{\mathrm{CS}} = \ln(N_c)/2$ as illustrated in Figure 4.9. Finally, we note that although CSMI, the upper-bound and their respective gradients can be expressed analytically, the gradient-based optimization is not guaranteed to converge toward a global optimum as the metric is not necessarily convex given the large number of classes and components.

4.4.2 Optimization Procedure

For the optimization of a projection operator P with the $\mathrm{P_e}$ upper bound depending on the CSMI metric, we can implement a simple gradient ascent algorithm : starting from an initial operator P_0 we update the coefficients of the matrix P_k at each step k with the

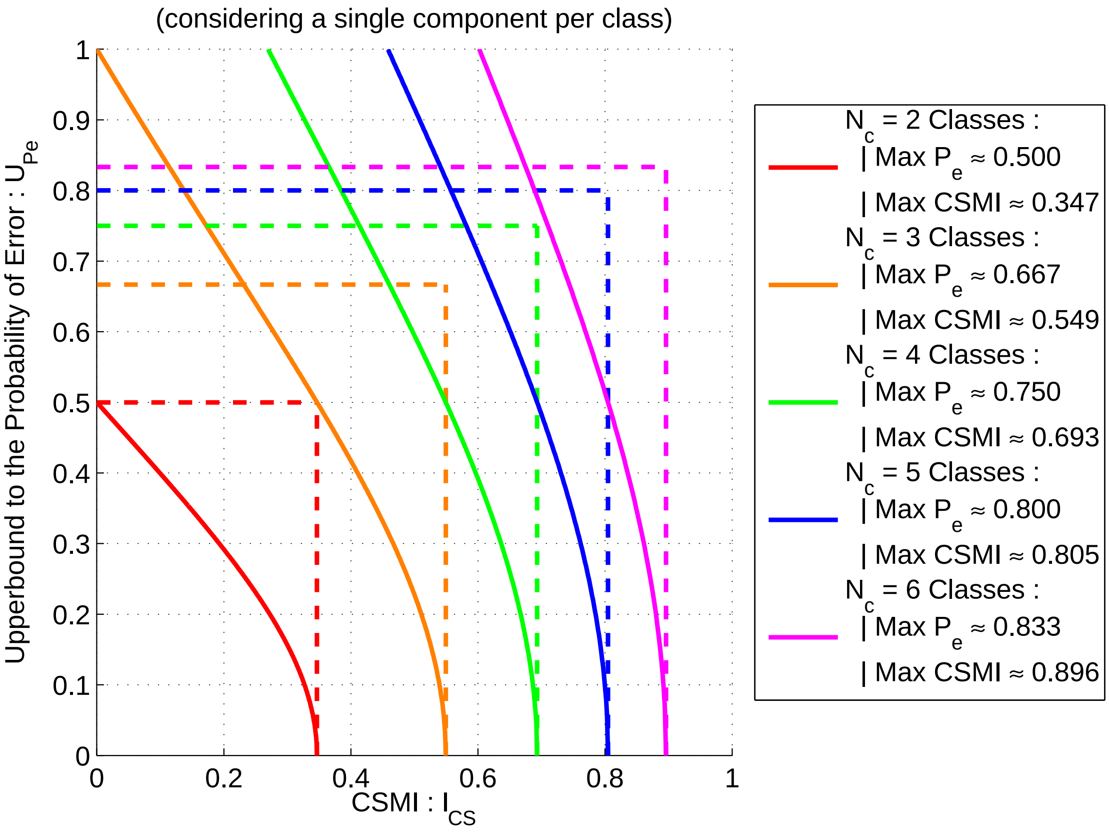


Figure 4.9: Upper bound $\mathrm{U_{Pe}}$ on the probability of error versus the CSMI metric $\mathcal{I}_{\mathrm{CS}}$ for different number of classes N_c. Here, we consider only a single component per mixture. The graph also shows the maximum values possible for both CSMI ($\mathcal{I}_{\mathrm{CS}} \leq \ln(N_c)/2$) and $\mathrm{P_e}$ ($\mathrm{P_e} \leq 1 - 1/N_c$).

gradient of $\mathrm{U_{Pe}}(P)$ (depending directly on the CSMI gradient $\boldsymbol{\nabla}\left(\mathcal{I}_{\mathrm{CS}}(P_k)\right)$ and given by Equation B.21 in section B.3 of the Appendix) : $P_{k+1} = P_k + \eta \boldsymbol{\nabla}\left(\mathrm{U_{Pe}}(P_k)\right)$, where η is an appropriate scalar step size. Note we also trim, at each step, the components of the gradient $\boldsymbol{\nabla}\left(\mathrm{U_{Pe}}(P_k)\right)$ which would cause P_{k+1} to step out of the physical operator space $[-1/M; 1/M]^{M\times N}$ justified previously. To implement this optimization algorithm we need to compute the CSMI metric and its gradient at each step of the optimization process.

From the previous expressions, one can verify that the complexity of the CSMI computation grows linearly with the dimensions of the direct space N and of the compressed space M, and quadratically with the total number of components across all classes $K = \sum_{i=1}^{N_c} O_i$, *i.e.* the computational complexity of CSMI is in the order of $\mathcal{O}(MNK^2)$. The computational complexity for the expression of the gradient follows a similar trend except for the quadratic increase with the number of dimensions of the direct space, *i.e.* the complexity is in the order of $\mathcal{O}(MN^2K^2)$. While the total computational cost is dominated by the latter, it is important to note that the search space is of dimension $M \times N$, that is several thousand dimensions in the current context, which is impractical for gradient-free optimization methods such as Simulated Annealing.

With the large mixtures being considered for this problem (on the order of one million components, each), the calculation of CSMI is computationally intractable on a single processor. It can naturally be split across multiple processors and benefits from a straightforward parallelization. Moreover, we also obtained satisfying results by implementing a procedure akin to the stochastic gradient ascent [71, 72] : at each step, the CSMI metric and its gradient are approximated by only computing the metric over a small subset of all the possible component pairs forming the previous potentials. In this subset we only include pairs of components which are narrowly separated. At each step k of the gradient ascent optimization, we select a new suitable subset at random and we compute the approximation of the metric $\mathcal{I}_{\text{CS}}{}_k^*$ and its gradient $\boldsymbol{\nabla}\left(\mathcal{I}_{\text{CS}}\right)_k^*$. To retain a short-term memory of the previous steps, we implement a simple decay function for the global estimate of CSMI $\mathcal{I}_{\text{CS}k}$ (and its gradient $\boldsymbol{\nabla}\left(\mathcal{I}_{\text{CS}}\right)_k$) as follow :

$$\begin{aligned} \mathcal{I}_{\text{CS}k} &= \left(\mathcal{I}_{\text{CS}}{}_k^* + \beta\, \mathcal{I}_{\text{CS}k-1}\right) && (4.14) \\ &= \left(\mathcal{I}_{\text{CS}}{}_k^* + \beta\, \mathcal{I}_{\text{CS}}{}_{k-1}^* + \beta^2\, \mathcal{I}_{\text{CS}}{}_{k-2}^* + \ldots\right) \\ \text{And similarly : } \boldsymbol{\nabla}\left(\mathcal{I}_{\text{CS}}\right)_k &= \left(\boldsymbol{\nabla}\left(\mathcal{I}_{\text{CS}}\right)_k^* + \beta \boldsymbol{\nabla}\left(\mathcal{I}_{\text{CS}}\right)_{k-1}\right) && (4.15) \end{aligned}$$

where the coefficient β is strictly smaller than one. Note that smaller β values are equivalent to faster *memory loss.*

The validity of the restriction to a particular subset can be illustrated by considering the distance separating the targets manifold : as the background accounts for a large portion of the chip signal, it is very likely that two targets with the same attitude but against different terrains are separated by a great $\mathcal{L}_2$ distance in the direct space. This remains true after the projection, most of the time. Hence, we can deduce that most of the Gaussian pairs component based on two different backgrounds do not contribute significantly to CSMI and their computation can be omitted. Each subset considered is therefore including targets (as means of the mixture components) sharing the same background. With this simplification, we obtained speedups of about two orders of magnitude when compared to the exhaustive approach. In addition, a software implementation tailored for the parallel architecture of Graphics Processing Units (GPUs) will be able to optimize a 24 × 1024 operator in about three hundred GPU-hours (on NVIDIA$^{\text{TM}}$ K20x), *e.g.* about 10 hours on 30 devices.

4.4.3 Optimization Results

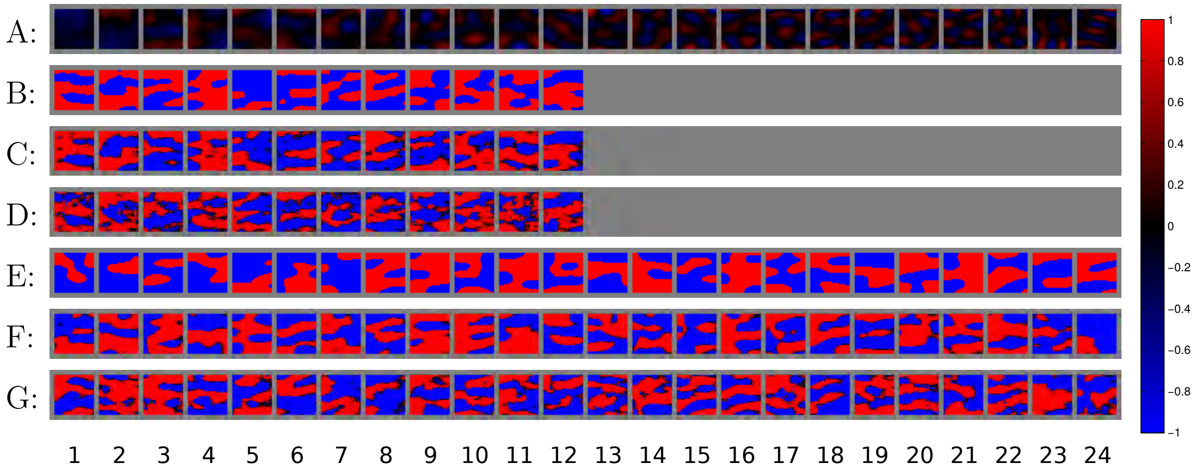


Figure 4.10: 32 × 32 modulation patterns. **A** : PCA-Secants, up to 24 projections ordered from left to right by decreasing prevalence; CSMI optimized operators, for $M = 12$ projections (**B,C,D**) and $M = 24$ projections (**E,F,G**). **B,E** are optimized for SNR=0dB, **C,F** are optimized for SNR=5dB; **D,G** are optimized for SNR=10dB. While all the modulation values are constrained to the $[-1/M, 1/M]$ range, all of the projections are here normalized to the $[-1, 1]$ interval for visualization.

The optimization procedure is repeated several time, for different number of projections, SNR values and initialized at different random Gaussian operators starting-points. A few of the patterns set obtained are shown in Figure 4.10 along with a PCA-Secants operator computed from the same data. As expected the optimized operators adopt different modulation strategies for different noise regimes. Firstly, the transmission is almost entirely saturated at low SNR when compared to PCAs. Secondly, the range of spatial frequencies addressed by each operator is increasing with SNR as the receding noise variance level uncovers more of the object Power Spectral Density (PSD). We can thus observe that most of the class information resides in the high spatial frequencies domain although it might not be interesting to perform measurements there when dealing with very noisy measurements.

Although less noticeable, one can remark that for high SNR, the operators with 12 projections present more gray-scale transmission values than those with 24 projections (similar to binary patterns). This can be justified as the formers are slightly information-starved while the latter are slightly energy-starved. This difference will be highlighted again with the results presented from the $\mathrm{P_e}$ performance curves. Finally, the optimized patterns exhibit a slight preference toward the vertical components of the targets spatial frequency spectrum : this is understandable as the vehicles in the current database are more horizontally elongated and therefore leak more energy along the vertical spatial frequencies.

4.5 System Performance Analysis

4.5.1 Simulation Study

To assess the performance of various operators, measurements are taken of numerous synthetic samples from the previously described realistic test set, as shown in Figure 4.6. They then serve as inputs to the staged classifier. This test dataset is entirely disjoint from the training samples and contains eight thousand chips, uniformly sampled across the two vehicles and the negative background classes. For the projection patterns, we include several common operators to be compared : signed random binary, random Gaussian, PCA-Secants (constructed from the secants between the means of the mixture components) and CSMI-Optimized patterns. For completeness, we also acquire measurements in the canonical space to simulate the performance of a standard imager using a complete exposure, *i.e.* the same flux as all the projections of a single operator. Those direct measurements are processed with the same two-stages classifier but which operates with the uncompressed statistical model as ancillary data rather than the compressed model.

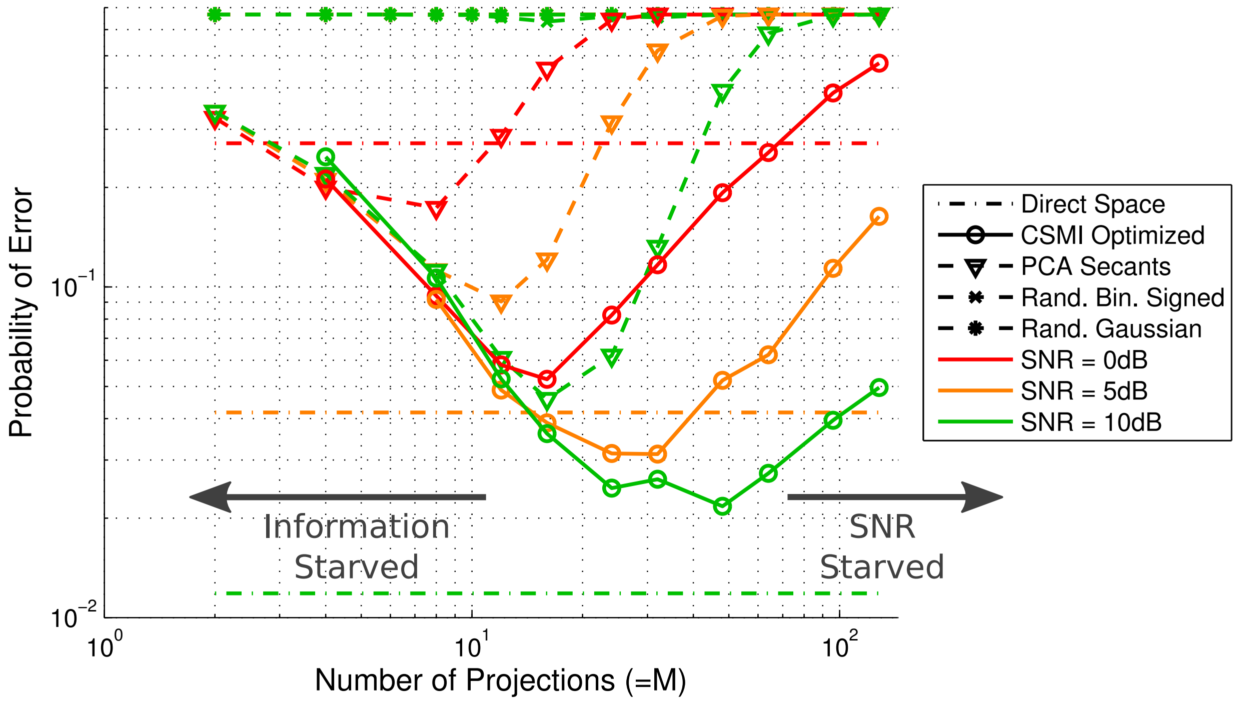


Figure 4.11: $\mathrm{P_e}$ performance in a simulated environment, averaged over the classification results of more than 8,000 realistic test samples. The horizontal lines indicates the performance of a direct imager and are independent of the number of projections M.

The results shown in Figure 4.11 demonstrates that optimum performances can be reached for compression rates between 64x and 42x (respectively between 16 and 24 projections for 32×32 blocks). For comparison, these are eight to ten times better than traditional rates for a compressive imaging application. At these compression rates, the probability of error for CSMI-Optimized operators obtained by simulation is between 3% (at 5dB) and 6% (at 0dB), and 1.7x to 3 times lower than for PCA-Secants.

In parallel, the performance gain (decreased $\mathrm{P_e}$) of the structured operators over random projections is clearly illustrated. Here, the energy normalization imposed to all these operators from the physical constraints of the system is indeed particularly detrimental to the random operators. This is because, their prior-less structure forces them to perform more projections in order to capture the image information, and thus the embedded class information, which is incompatible with the finite flux constraint. This normalization is also the reason for the valley-shaped curves : the operators cannot successfully acquire the target information from a small number of patterns projections (information-starved), while they inefficiently spread the incoming flux across too many inherently-noisy measurements for a large number of patterns projections (SNR-starved).

Finally, the $\mathrm{P_e}$ curves also highlight the benefit of compressive sensing : a significant multiplexing gain over direct measurements is visible at low SNR with an error rate 5 times smaller at 0dB. At 10dB however, we notice that the performance of the Information Optimal projections does not reach that of a conventional imager. This can be partially imputed to a failure of the optimization framework : as the SNR increases, the blur imposed to the statistical model by the noise kernel is reduced and it reveals more complex regions in the CSMI surface. These particular regions are non-convex and cannot be efficiently explored solely with a traditional gradient ascent approach.

4.5.2 Experimental Validation

For the experimental qualification of the operators, we perform all the measurements with our prototype described in chapter 3 and pictured in Figure 4.12. For this acquisition, each test sample is shown on a screen facing the main objective and imaged onto a Digital Micromirror Device (DMD). After the optical modulation, the signal is integrated by a custom relay optics onto a low-resolution CCD sensor and the measurements are transferred to a computer for processing. In order to minimize distortions between the ideal measurement model presented in Equation 4.2 and the physical measurement model implemented, an automated calibration is performed once before the acquisition session, as described in the previous chapter. It consists in analyzing the mapping between the 2×2 blocks of 32×32 micromirrors (or modulation elements) and the pixels of the sensor array. The bias and linearity of the measurements are also evaluated by displaying a few uniform grayscale images on the screen. The experiment tests set is composed of one thousand 64×64 pixels chips randomly selected from the realistic tests dataset. Before being shown, they are pre-corrected to neutralize the dynamic range compression applied by the display panel (with a constant $\gamma \approx 2.2$). Finally, for each of the operators to be tested, we acquire and average ten consecutive measurements from the prototype in order to reach a SNR exceeding 30dB. This way, we can add simulated white Gaussian noise of a given variance in order to adjust the measurement SNR before the classification, and up to 20dB at most. For this last step, we use the same algorithm as for the simulation study and we repeat the process with several noise realizations.

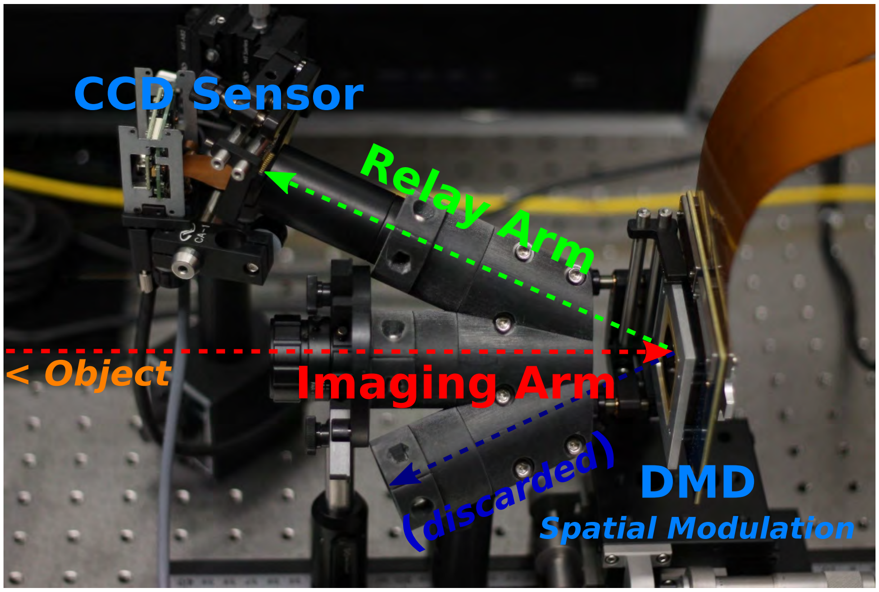


Figure 4.12: Structure of the Scalable Compressive Camera prototype used to acquire the experimental data. The scenes are shown onto a display at the left of the imaging arm. The modulation patterns are implemented successively by a DMD and the resulting signal is integrated by the relay arm onto a low resolution CCD sensor.

The experimental results are reported in Figure 4.13 for 12 and 24 projections operators, along with the corresponding entirely simulated measurements for reference. We can remark analogous trends between the the experiment and the simulation with varying SNR values. Especially, one can notice that for $M = 24$, the operator optimized for SNR = 0dB outperform the operator optimized for SNR = 10dB until SNR ≈ 5dB (≈ 6dB from the simulation). For $M = 12$, all the operators are in an information-starved regime : the number of projections performed is not sufficient to accurately discriminate the class label. As demonstrated in the previous simulation analysis, the PCA-Secants operator underperforms the Information Optimal patterns in this low SNR regime, with P_e improvement factors between 2x at 10dB, and 6x at 0dB for 24 projections.

Despite the careful calibration, we note that it is however not possible to exactly reach the performance predicted by simulations as illustrated by the previous averaged results : we find an increase between 2 and 5% of the probability of error for the tested operators. Those differences are mainly caused to several imperfections in the experimental acquisition setup. For instance, the display used does not reproduce the linear intensity controls with a high fidelity : hence the targets shown are still suffering from a slight dynamic range compression. The PSF of the imaging arm will also slightly blur some of the smallest scene details.

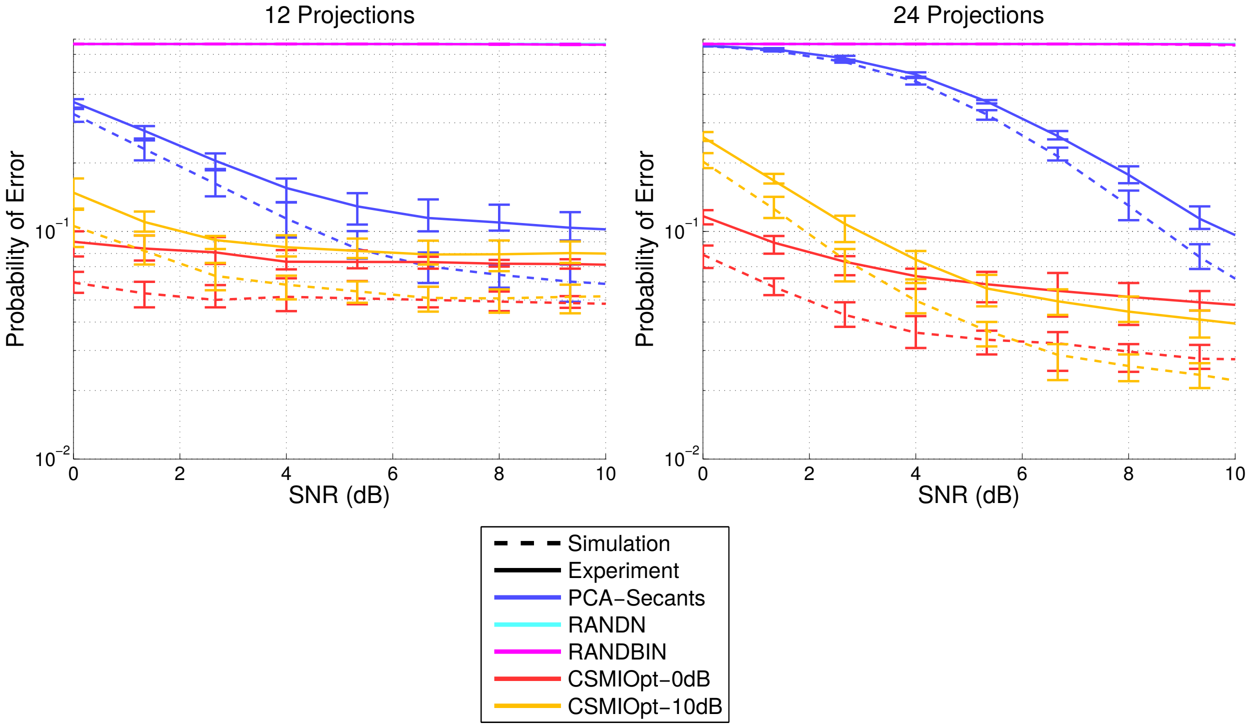


Figure 4.13: Comparison of between the simulated and experimental performance averaged over various realization of the noise. The error bars indicate the minimum and maximum P_e found across these realizations.

4.6 Conclusion And Future Work

We have presented a scalable optical compressive sensing architecture to perform an Automated Target Detection and Recognition task. This platform optically decomposes the field of view into contiguous blocks and acquire for each a feature vector that an algorithm then classifies without performing any intermediate image reconstruction. This piecewise method allows for the parallel acquisition and processing of the measurements taken with a low resolution sensor. In the context of this study, we have developed a large statistical model of the target vehicles within cluttered natural environments, and with it we optimized the projection patterns of the sensing matrix by maximizing the Cauchy-Schwarz mutual information which measures the separation between the classes after projection. We have especially demonstrated that this metric and its gradient can be computed in closed-form even for our complex statistical mixture model, and that it provides an upper bound on the overall probability of misclassification of the system. The projection operators thus constructed are tuned for a fixed number of projections and acquisition SNR. We analyze their performances and demonstrated that they outperform conventional cameras as well as random and PCA-secants operators in both simulated and experimental setups, reaching accuracy rates of about 90% at 0dB SNR above 95% at 10dB SNR.

In future developments, we plan on adding to the complexity of the statistical models in order to handle, for instance, occlusions and change in lighting conditions. We can also now add sensing capabilities to the system, such as estimating the motion parameters and tracking the targets. Finally with a programmable compressive architecture, it is possible to switch on demand between an autonomous target detection and recognition task described here and an imaging task described in the previous chapter, for only a few marked regions of interest.

Chapter 5

Two Point-Sources Resolution Measurement Design and Analysis

In this final chapter, we consider the canonical problem of estimating the angular separation between two incoherent and quasi-monochromatic point-sources in order to investigate the resolution limit of optical instruments. We expand on recent developments showing that there exists a passive and linear optical architecture which outperforms the conventional direct imager in the sub-Rayleigh regime. We especially provide a generalization of this two-measurements architecture to match arbitrary pupil functions and show that it remains insensitive to diffraction and phase aberrations.

5.1 Introduction

The Rayleigh Criterion [73] provides an intuitive qualitative limit to the resolution of a traditional diffraction-limited imager employing a focal plane array, as depicted in Figure 5.1. It reasonably states that two incoherent point-sources are resolvable if their separation is greater than the extent of the Point Spread Function (PSF) due to the diffraction from the finite size of the pupil. Consequently, attempting to estimate smaller separation angles is infeasible because the smeared images of the points in the observation plane overlap with each other rendering the two point-sources visually indistinguishable. Increasing the resolution of the focal plane does not fundamentally change the resolution as this limit appears to derive solely from the physical angular size of the PSF which is in the order of λ/D for a λ-wavelength field passing through a pupil of scale D. Modern techniques have been developed to overcome this resolution limit such as near-field scanning microscope, stimulated emission depletion (STED) microscopy [79, 80] or photoactivated localization microscopy [81]. However, these methods cannot be adapted to long standoff imaging or remote sensing applications as they require either close proximity to exploit the short distance propagation of evanescent waves or elaborate control of the object illumination to restrict the excitation pattern of fluorescent markers.

However in recent work by Tsang *et al.* [83, 84] have shown that the common perception of Rayleigh's curse is merely a limitation of the direct sampling strategy that conventional imagers employ rather than a fundamental physical limit of all instruments. To demonstrate their argument, they have computed the quantum Fisher information (QFI), and the associated quantum Cramer-Rao lower bound (QCRLB) to find the ultimate achievable precision of any optical measurement for the task of measuring the angular separation between two quasi-monochromatic incoherent point-sources. Contrary to the performance of a traditional imaging architecture, this information bound remains constant throughout the whole range of angular separations, *i.e.* from zero to large separations in the image plane. More surprisingly, the authors have proposed a passive linear optical system which attains this bound and thus outperforms any conventional imager given an optical aperture shape. In this alternate architecture, depicted in Figure 5.2, the focal plane array is replaced with an optical preprocessing device which performs the decomposition of the incident diffracted fields onto the Hermite-Gauss spatial modes and counts the number of photons in each mode using shot-noise limited photodetectors. The linear projections onto the mode basis allows extract the phase information embedded in the optical fields, while this same phase information is not captured by the traditional focal plane array measurements. The measurements quantum optimality was further investigated by Lupo [85].

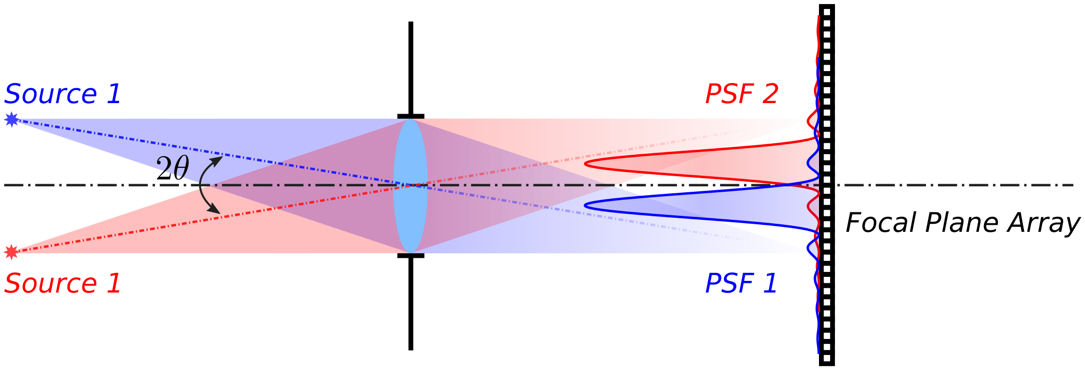


Figure 5.1: Direct imaging architecture : the focal plane array records the sum of the intensity of the two PSFs.

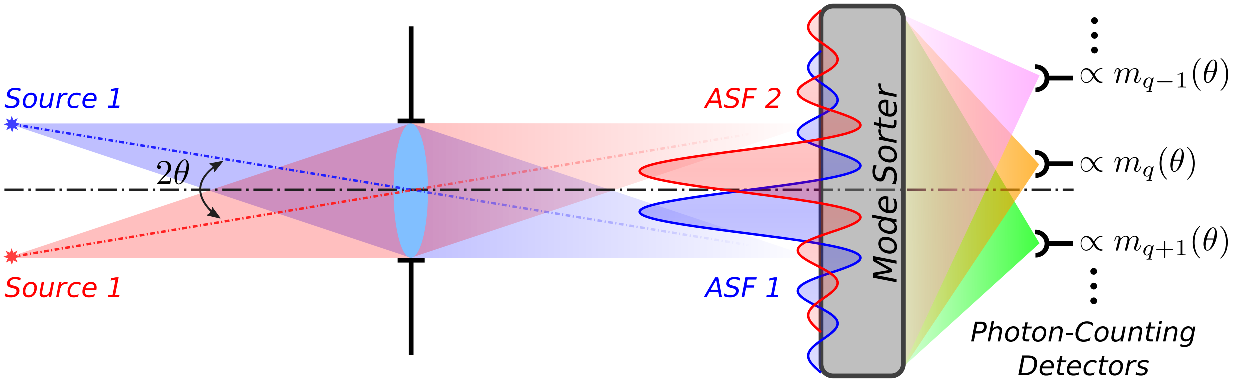

Source 1
2θ
Source 1
ASF 2
ASF 1
Mode Sorter
$\propto m_{q-1}(\theta)$
$\propto m_q(\theta)$
$\propto m_{q+1}(\theta)$
Photon-Counting Detectors

Figure 5.2: Mode-sorting measurement architecture : a passive optical element linearly decomposes the incident fields onto a spatial modes basis. For each mode, a shot-noise limited photodetector integrates the flux at the output port.

In this work, we extend Tsang's *et al.* architecture for a Gaussian aperture first to a hard aperture, before generalizing to an arbitrary pupil function. We explore several key properties of the proposed non-traditional system : (1) energy and (2) information distribution across the optical modes via an analysis of the classical Fisher information and (3) the mean square error performance of the Maximum Likelihood Estimator (MLE). We especially analyze aspects of this novel imaging system that will impact its performance subject to non-idealities inevitable in a practical implementation of the system.

This chapter is organized as follow. In section 5.2, we begin by reviewing the measurement statistics governing the angular accuracy of a direct, or conventional, imager employing an idealized detector array with infinite resolution. The Fisher information and the estimation error for this conventional imager will serve as a baseline for comparison with alternative architectures. We then describe the measurements strategy of the spatial mode demultiplexer for Gaussian-apodized and hard apertures in section 5.3. We show that the Sinc-Bessel spatial modes basis is the optimal measurement for a hard aperture. We highlight the duality between energy and information content of the spatial modes for the mode-sorting architecture and demonstrate that a binary mode measurement can outperform a conventional imager recording thousands of incident photons in the sub-Rayleigh range. In section 5.4, we complement the Fisher information analysis of the normalized root mean square error (NRMSE) performance of measurements by the mode-sorter and the conventional imager for different source fluxes. In section 5.5, we generalize the mode-sorting architecture to an arbitrary aperture and we highlight its advantages over conventional imaging. Finally in section 5.6, we discuss two candidate implementations of the mode-sorting measurements : (1) a Mach-Zehnder interferometer architecture and (2) a self-referencing volume hologram architecture.

5.2 Fundamental Performance Analysis of A Conventional Imager

While Lord Rayleigh's criterion describes a qualitative visual limit to the resolution of an incoherent imaging system, we develop a quantitative analysis of its fundamental performance limit via the Fisher information. Such a quantitative analysis is essential to establish a baseline for comparison of alternate imaging architectures. In this one-dimensional problem, two incoherent, indistinguishable and quasi-monochromatic (λ) point-sources are considered, each respectively at an angle $\pm\theta$ from the optical axis, small enough to permit the use of the scalar diffraction theory. A conventional imager with an aperture scale D and a linear shift-invariant impulse response, maps the incident optical field onto its focal plane array as the superposition of two shifted Amplitude Spread Function (ASF) : $A_\sigma(x) = A(x/\sigma)/\sqrt{\sigma}$ where x is an angular coordinate and σ is a normalization constant proportional to λ/D. As dictated by the scalar theory of wave optics, this characteristic function (ASF) of the system is directly proportional, in amplitude and physical size, to the Fourier transform of the complex aperture function. However in our current analysis only the intensity point spread function (PSF) is of interest for incoherent imaging. The PSF is given by the intensity of the ASF : $P_\sigma(x) = |A_\sigma(x)|^2$. Note that throughout the remainder of this work, both the ASFs and the PSFs will refer to the normalized version in energy : *i.e.* $\int_{-\infty}^{\infty} P_\sigma(x)\ dx = 1$. For example, a Gaussian-apodized aperture $T(x') = \exp\left(x'^2/2D^2\right)$ yields a Gaussian ASF $A_\sigma(x) = \exp(-x^2/4\sigma^2)/(2\pi\sigma^2)^{\frac{1}{4}}$, and a PSF $P_\sigma(x) = \exp(-x^2/2\sigma^2)/\sqrt{2\pi}\sigma$. For a hard aperture of size D, *i.e.* $T(x') = \mathrm{rect}(x'/D)$, the ASF is given by the cardinal-sine function $A_\sigma(x) = \mathrm{sinc}(x/\sigma)/\sqrt{\sigma}$, whereas the PSF can be expressed as $P_\sigma(x) = \mathrm{sinc}(x/\sigma)^2/\sigma$.

For the task at hand, we assume that only the 2θ-separated point-sources compose the field of view, each emitting a flux of $N/2$ photons on average, incident on the exit pupil. Therefore, the incoherent image in the focal plane of the imager can be written as $I_\sigma(x,\theta) = N(P_\sigma(x+\theta) + P_\sigma(x-\theta))/2$. Naturally, this irradiance function describes the average distribution of photons over the observation plane and solely depend on the separation angle θ. Given a thermal model of object radiance, an ideal shot-noise limited photodetector of size ϵ centered at the coordinate x_0 with 100% quantum efficiency measures a random flux W that follows a Poisson statistical distribution with mean parameter equal to the integrated irradiance above its surface : *i.e.* $W \sim \mathcal{P}\left(\Lambda_\epsilon = \int_{x_0-\epsilon/2}^{x_0+\epsilon/2} I_\sigma(x,\theta)\ dx\right)$. Note that a contiguous tiling of such pixels forms an infinite sensor array which collects a random total number of photon W_t which also follows a Poisson distribution with mean N, that is :

$$\begin{aligned} W_t &\sim \sum_{k=-\infty}^{+\infty} \mathcal{P}\left(\int_{k\epsilon-\epsilon/2}^{k\epsilon+\epsilon/2} I_\sigma(x,\theta)\ dx\right) \\ &\sim \mathcal{P}\left(\sum_{k=-\infty}^{+\infty} \int_{k\epsilon-\epsilon/2}^{k\epsilon+\epsilon/2} I_\sigma(x,\theta)\ dx\right) \sim \mathcal{P}\left(N\right). \end{aligned} \tag{5.1}$$

Moreover, in the limit $\epsilon \to 0$, the detector array approaches a continuum detector such that flux statistics for each of the infinitely small detector tends to a Bernoulli distribution of probability $\Lambda_{\epsilon,k} = \int_{k\epsilon-\epsilon/2}^{k\epsilon+\epsilon/2} I_\sigma(x,\theta)\, dx \to 0$. Hence, the distribution of the photon counts for such a continuum spatial detector is defined by an inhomogeneous Poisson point process with density given by the irradiance function $I_\sigma(x,\theta)$. In other words, such a detector measures a random number $(W_t \sim \mathcal{P}(N))$ of photons during the exposure, where each photons detection is independent of each other and has a probability that it is detected in the interval $[x_a, x_b]$ given by $\int_{x_a}^{x_b} I_\sigma(x,\theta)\, dx$ as symbolically illustrated in Figures 5.3 and 5.4.

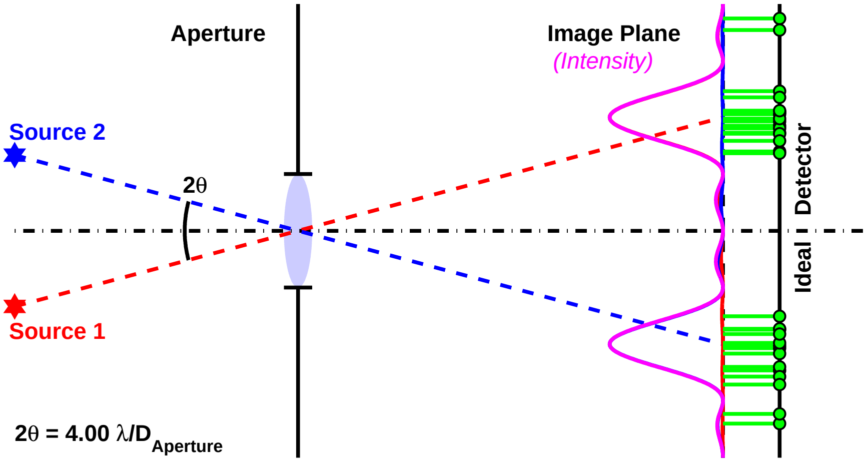


Figure 5.3: Direct imaging architecture for two incoherent and well separated point-sources. Green dots depict the photon detection events.

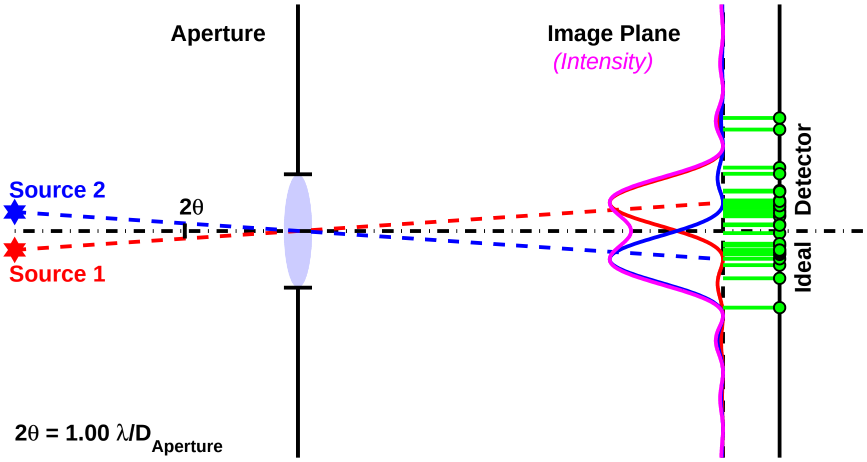


Figure 5.4: Direct imaging architecture for two incoherent point-sources at Rayleigh's limit. Green dots depict the photon detection events.

Therefore, the likelihood $\mathcal{L}_{\text{Direct}}$ of detecting N^* independent photons $\{x_1, \ldots, x_{N^*}\}$ can be expressed as the product of the irradiance densities at each detected location :

$$\mathcal{L}_{\text{Direct}}(x_1, \ldots, x_{N^*}|\theta, N^*) = \prod_{k=1}^{N^*} I_\sigma(x_k|\theta). \tag{5.2}$$

Given this likelihood model, one can calculate the Fisher information $\mathcal{I}_{\text{Direct}}$ [74] for estimating the variable θ (angular separation) which can be expressed as :

$$\begin{aligned}\mathcal{I}_{\text{Direct}}(\theta|N^*) &= \mathop{\mathbf{E}}_{x_1,\ldots,x_{N^*}}\left[\left(\frac{\partial\,\mathcal{L}_{\text{Direct}}(x_1, \ldots, x_{N^*}|\theta, N^*)}{\partial\theta}\right)^2 \middle|\, \theta\right] \\ &= N^* \int_{-\infty}^{+\infty} \frac{I'(x,\theta)^2}{I(x,\theta)}\, dx, \end{aligned} \tag{5.3}$$

where $I'(x,\theta)$ denotes the first derivative of the irradiance function with respect to θ, *i.e.* $I'(x,\theta) = \partial I(x,\theta)/\partial\theta$; and thus on average :

$$\mathcal{I}_{\text{Direct}}(\theta) = \mathop{\mathbf{E}}_{N^*}\left[\mathcal{I}_{\text{Direct}}(\theta|N^*)\right] = N \int_{-\infty}^{+\infty} \frac{I'_\sigma(x,\theta)^2}{I_\sigma(x,\theta)}\, dx, \tag{5.4}$$

This $\theta-$differential FI metric assigns a score of the system performance for a given separation (θ) : higher values of the metric indicate better performances. One can observe several interesting properties of the FI expression in Equation 5.4 and as plotted in Figure 5.5 as a function of point-sources separation. Firstly, the FI is linearly dependent on the number of photons. Secondly, the FI value also approaches zero when the separation angle θ tends to zero, and regardless of the underlying PSF profile (see Proof 4 in the Appendix). This is consistent with Rayleigh's criterion and establishes the insensitivity of any direct measurement device to angular separations below the Rayleigh limit. Thirdly, the amount of information depends on the pupil function. For instance an imaging system with a hard aperture (*i.e.* $T(x') = \text{rect}(x'/D)$) collects more information than a system with a Gaussian aperture. Moreover, any optical aberrations in the exit pupil of the optics increases the PSF blur and consequently lower $\mathcal{I}_{\text{Direct}}$ for all values of the separation θ. Finally, FI saturates close to a maximum value when θ becomes much larger than the physical extent of the PSF.

This is to be expected as the performance of the imaging system is nearly independent of the separation when the two source images are easily distinguishable (*i.e.* well above Rayleigh limit).

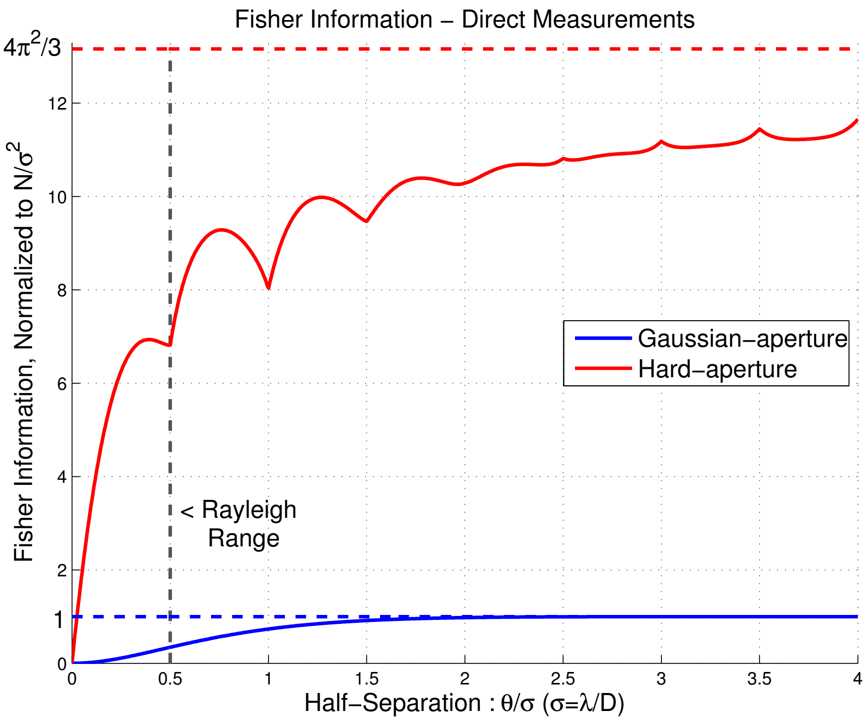


Figure 5.5: Fisher information of idealized direct imagers for Gaussian and hard apertures. Note that the *y-axis* is normalized to the average number of photons N and divided by the square of the PSF scale σ. The ringing in the hard aperture curve is due to the features in the sinc² PSF.

The Fisher information also provides a lower bound to the mean square error (MSE) of any unbiased estimator also known as the Cramer-Rao Lower Bound (CRLB) [75, 76]. Thus for an estimator $\widehat{\theta}(x_1, \ldots)$ using the detected photon locations $\{x_1, \ldots\}$, we can state : $\mathrm{MSE}_{\widehat{\theta}}(\theta) = \mathbf{E}\left[(\widehat{\theta}(x_1, \ldots) - \theta)^2 \middle| \theta\right] \leq 1/\mathcal{I}(\theta)$. However, it is important to emphasize that this bound is not necessarily achievable. In order to test the direct imager output performance, we first construct the maximum likelihood estimator (MLE) $\widehat{\theta}_{\text{MLE,Direct}}(x_1, \ldots, x_{N^*})$ from Equation 5.2 as :

$$\widehat{\theta}_{\text{MLE,Direct}}(x_1, \ldots, x_{N^*}) = \arg\max_{\theta} \left\{\mathcal{L}_{\text{Direct}}(x_1, \ldots, x_{N^*}|\theta)\right\}, \tag{5.5}$$

and then we compare its MSE performance against the CRLB. However for this angular estimation task, it is more appropriate to consider a normalized version of the MSE defined as the normalized root mean square error (NRMSE) as : $\mathrm{NRMSE}_{\widehat{\theta}}(\theta) = \sqrt{\mathrm{MSE}_{\widehat{\theta}}(\theta)}/\theta$, and equivalently the normalized Cramer-Rao lower bound (NCRLB) as : $\mathrm{NCRLB}(\theta) = \mathrm{CRLB}(\theta)/\theta$.

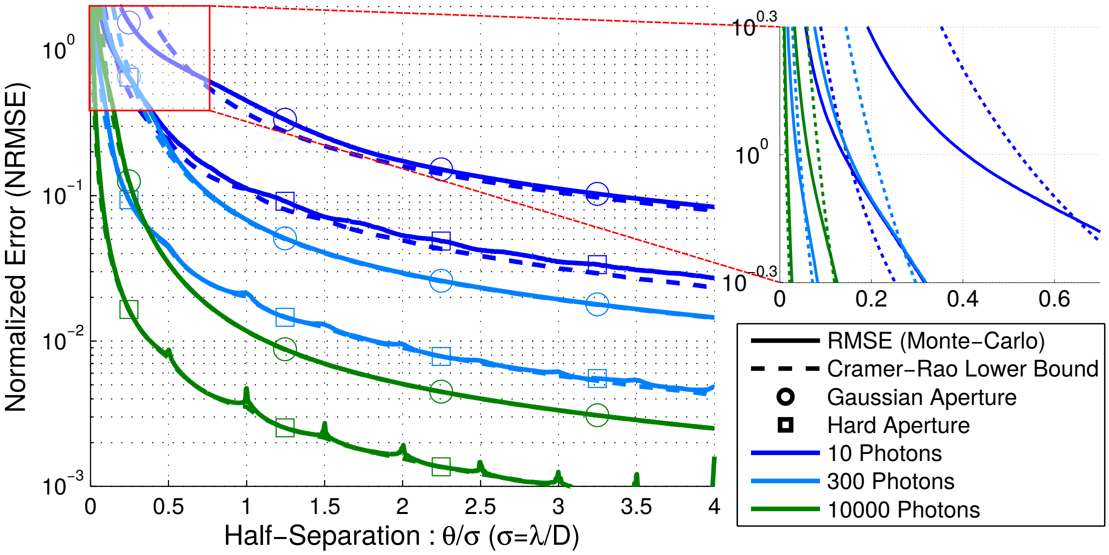


Figure 5.6: Normalized error plot ($\mathrm{NRMSE}_{\mathrm{Direct}}$) for idealized direct imagers for Gaussian and hard apertures. The normalized CRLB ($\mathrm{NCRLB}_{\mathrm{Direct}}$) is also included for reference.

To implement a near optimal MLE estimator performance we use Nelder-Mead search algorithm [77] initialized in the close proximity of the ground-truth value of θ so as to avoid possible local maxima. In Figure 5.6, we show the estimation performance of the ML estimator obtained using a Monte Carlo sampling of the NRMSE metric. The ML estimator is exposed over more than 16,000 sets of photons realizations for every sample value of θ and average incident flux N. From this plot we can observe that the NRMSE performance of the ML estimator is nearly in perfect agreement with the corresponding NCRLB for large mean-number of photons. However the MLE breaches the NCRLB for small number of photons ($\approx$ 10) and separations ($\leq$ 0.5), as it becomes biased (and super-efficient) in that regime. Finally, while the overall estimation performance increases with the square-root of the number of photons, the vertical asymptote in $\theta = 0$ indicates the inherent inefficiency described previously. For instance, in order to perform a measurement with an relative precision of 1% at a separation of about $4\lambda/D$ (4 times Rayleigh's limit), only 300 photons are needed on average for an imaging system with a hard aperture. However, this number is increased to 10,000 photons to achieve the same relative precision at a separation of about $0.8\lambda/D$ (20% smaller than Rayleigh's limit). In other words, reducing the angular separation by a factor of 5 requires an increase in the number of photon budget by over a factor of 30.

5.3 Analysis of Mode-Sorting Imager Design : Gaussian and Hard Apertures

An alternate approach to measuring the angular separation of two quasi-monochromatic incoherent point-sources proposed by Tsang *et al.* in [83, 84] involves projecting the superposition of the incoherent image optical fields $A_\sigma(x+\theta)$ and $A_\sigma(x-\theta)$ of the two point-sources onto a $\mathcal{L}_2$ complex-valued spatial mode $f(x)$ of unit magnitude, *i.e.* $\int_{-\infty}^{\infty} |f(x)|^2 \; dx = 1$. The resulting measurement function $m_f(\theta)$, parametrized by the separation angle θ can be analytically written as :

$$m_f(\theta) = \frac{1}{2}\left|\int_{-\infty}^{+\infty} \overline{f(x)} A_\sigma(x+\theta) \; dx\right|^2 + \frac{1}{2}\left|\int_{-\infty}^{+\infty} \overline{f(x)} A_\sigma(x-\theta) \; dx\right|^2 , \tag{5.6}$$

where $\overline{f(x)}$ denotes the mode complex conjugate. The normalized output of this measurement process is thus a single scalar equal to the sum of the two modulated intensities as defined in Equation 5.6 and is illustrated in Figures 5.7 and 5.8. Furthermore, this measurement strategy can be extended to multiple mode measurements simultaneously, which would lead to a linear decomposition of the image plane optical field onto any set of modes forming an orthonormal basis.

5.3.1 Energy Distribution Across Mode Basis

Given such a mode decomposition measurement architecture one can effectively count the number of photons populating an arbitrary spatial mode $f(x)$. Thus, if N denotes the total photon flux through the exit pupil, a single shot-noise limited bucket photodetector will measure samples from a Poisson distribution with mean $Nm_f(\theta)$. If multiple mode measurements are acquired, note that their statistics remain independent because the underlying orthogonal spatial modes do not interact.

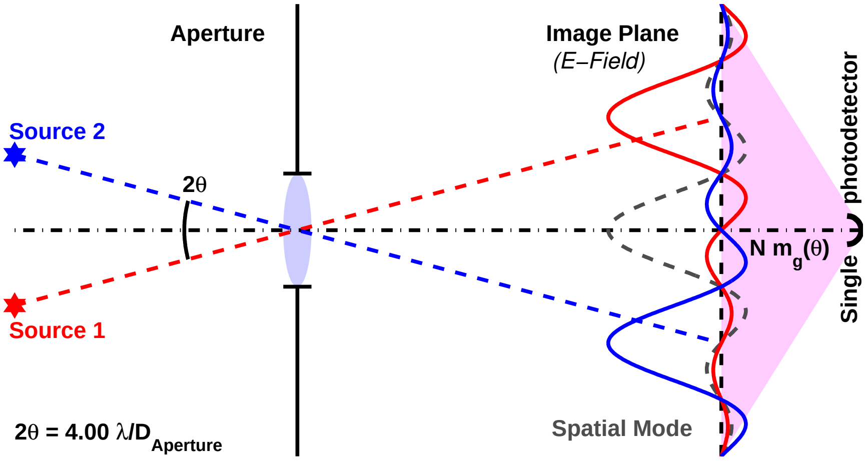


Figure 5.7: Single mode-sorting architecture : well separated point-sources. The fields are not coherent and cannot be summed.

In [83], the authors propose to use the Hermite-Gauss (HG) basis to decompose the Gaussian ASFs deriving from a Gaussian apodized aperture. For this particular case, $\{h_q(x)\}$ denotes the spatial mode set indexed by q, and $\{m_q(\theta)\}$ denote the associated normalized measurement functions defined in Equation 5.6. Mathematically, the mode $h_q(x)$ and the corresponding mode measurement $m_q(\theta)$ can be expressed as :

$$\forall q \in \mathbb{N},\ h_q(x) = \frac{(2\pi\sigma^2)^{-\frac{1}{4}}}{\sqrt{2^q q!}} H_q\left(\frac{x}{\sqrt{2}\sigma}\right) \exp\left(-\frac{x^2}{4\sigma^2}\right) \tag{5.7}$$

$$m_q(\theta) = \frac{1}{q!}\left(\frac{\theta^2}{4\sigma^2}\right)^q \exp\left(-\frac{\theta^2}{4\sigma^2}\right) \tag{5.8}$$

where $H_q(x)$ denotes the q^{th} Hermite polynomial and $h_0(x)$ is simply a Gaussian mode

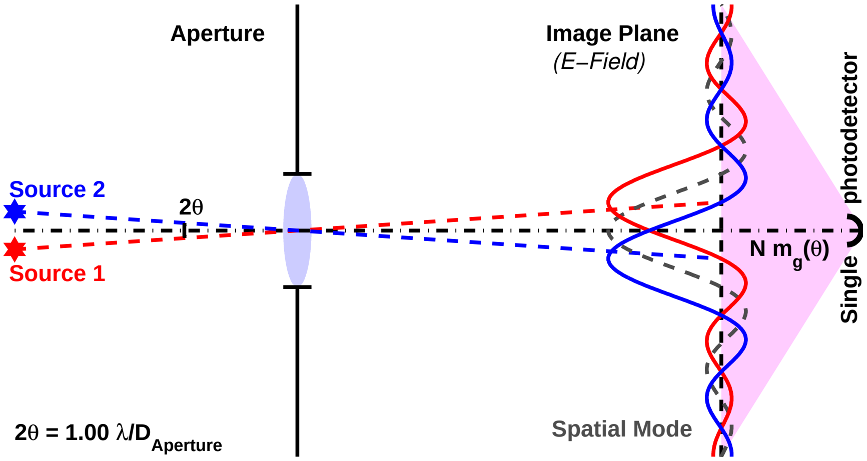


Figure 5.8: Single mode-sorting architecture : point-sources at Rayleigh's limit. The fields are not coherent and cannot be summed.

identical to the Gaussian ASF of the imaging system. For a system with a hard aperture, we find that the Sinc-Bessel (SB) orthonormal basis $\{s_q(x)\}$ is the optimal choice for the mode decomposition of the cardinal-sine ASFs. Thus for the hard aperture, we can express the q^{th} mode $s_q(x)$ and the corresponding mode measurement as :

$$\forall q \in \mathbb{N},\ \ s_q(x) = \sqrt{1+2q}\ j_q(\pi x) \tag{5.9}$$

$$m_q(\theta) = s_q\left(\frac{\theta}{\sigma}\right)^2 \tag{5.10}$$

where $j_q(x)$ is the q^{th} spherical Bessel function of the first kind (note especially that $j_0(\pi x) = \operatorname{sinc}(x)$). The HG and SB modes are plotted in Figure 5.9. Note that both of the proposed basis are also complete in the $\mathcal{L}_2$ vector space and consequently, the Parseval-Plancherel theorem ensures the conservation of energy in the measurements space. In other words, one we can assert that measuring all of the modes capture all of the incident energy, mathematically :

$$\forall \theta \in \mathbb{R}^+,\ \left\langle \sum_{q=0}^{+\infty} \mathcal{P}\left(N m_q(\theta)\right) \right\rangle = N \sum_{q=0}^{+\infty} m_q(\theta) = N. \tag{5.11}$$

Another key aspect shared by the two modes basis is that the projection onto the 0^{th}

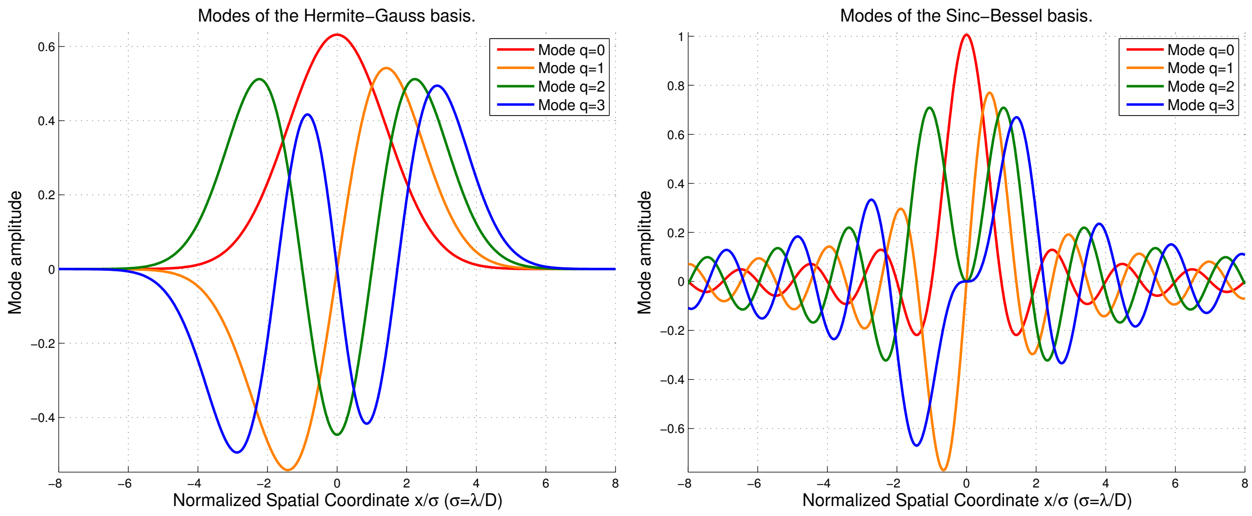


Figure 5.9: **Left** : first modes of the Hermite-Gauss basis. **Right** : first modes of the Sinc-Bessel basis.

mode captures all of the energy available at $\theta = 0$: *i.e.* $m_0(0) = 1$, and all the other modes capture, of course, no photons at zero separation. This is illustrated in Figure 5.10, where we plot the cumulative energy collection by measuring the modes 0 to Q. Note also that the incident energy from the sources is almost entirely captured by low order modes at small angular separations : for instance, in the case of the HG modes and a Gaussian-apodized aperture, cumulating the measurements of the modes $q = 0$ and $q = 1$ is sufficient to acquire more than 99% of the available energy at $\theta = \sigma/2$. In the case of the SB modes and a hard aperture, cumulating the measurements of the modes $q = 0, 1$ and 2 is sufficient to acquire more than 97% of the available energy at $\theta = \sigma/2$.

5.3.2 Fisher information Distribution Across Modes Basis

As we have shown, all the spatial mode measurements are statistically independent and follow shot-noise statistics with mean $m_q(\theta)$. If we begin with a single mode measurement denoted by $m_f(\theta)$ for a total average photon flux N, then the measurement likelihood of the associated measurement variable Y_f is given simply by a Poisson distribution conditioned on the angle θ : $Y_f \sim \mathcal{P}\left(N m_f(\theta)\right)$. For such a measurement we can express the corresponding

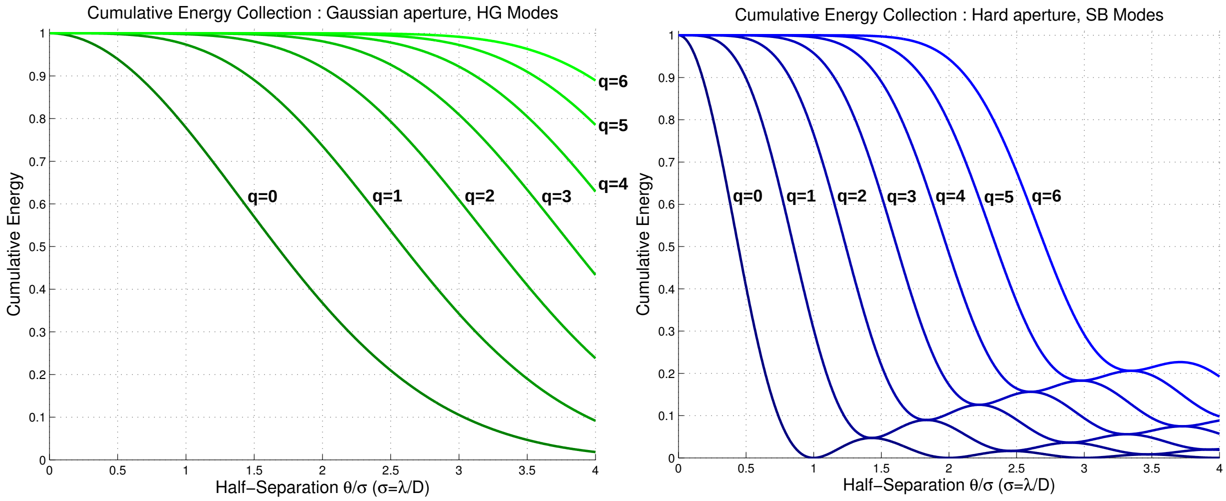


Figure 5.10: Fraction of the energy collected over the cumulative measurements $0, \ldots, q$. **Left** : for a Gaussian aperture and HG modes. **Right** : for a hard aperture and SB modes. The lowest order modes allow to integrate most of the energy available at small separation angles θ.

Fisher information as :

$$\mathcal{I}_f(\theta) = N \frac{m_f'(\theta)^2}{m_f(\theta)}, \tag{5.12}$$

where $m_f'(\theta)$ denotes the first derivative of m_f with respect to θ : $m_f'(\theta) = \partial m_f(\theta)/\partial\theta$ (see Proof 1 for the derivation of Equation 5.12). It is interesting to note the underlying structure of the expression which is proportional to the square of the measurement sensitivity to θ (*i.e.* $m_f'(\theta)$) and inversely proportional to the measurement itself ($m_f(\theta)$). In particular, the FI structure matches that of the Fisher information for a scalar measurement function $g(\theta)$ corrupted by an additive Gaussian noise of variance ν^2 : $g'(\theta)^2/\nu^2$.

Despite this analogy, we must highlight that the variance of the Poisson distribution is signal dependent, which is not the case for Gaussian noise corrupted measurement. The FI of the Poisson distributed mode measurement exhibits a counter-intuitive property : if the measurement function approaches 0 (*i.e.* no energy is collected), while its sensitivity also approaches 0 near a critical point θ_c, furthermore if the measurement function is equivalent to the square of its sensitivity, then the Fisher information itself can have a non zero value at θ_c, mathematically :

$$\exists C > 0, \quad C m_f(\theta) \underset{\theta \to \theta_c^{\pm}}{\sim} m_f'(\theta)^2 \quad \Rightarrow \quad \mathcal{I}_f(\theta) \underset{\theta \to \theta_c^{\pm}}{\to} C \tag{5.13}$$

Such a measurement property suggests that every single detected photon, for an unknown parameter θ in the vicinity of θ_c, carries a lot of information about the parameter and thus mitigates its decaying sensitivity $(m_f'(\theta))$. Interpreting this result for the two point-sources separation problem suggests that if a mode measurement has this inherent property (Equation 5.13) then it provides some information around θ_c, even though its sensitivity is near zero and it collects near zero energy. However, an important practical implication of this property is that around the critical point θ_c the information collection represent the equivalent of an unstable mechanical equilibrium, that is any corruption of the measurement, such as a background signal, will immediately degrade the Fisher information content to zero around θ_c. Thus, such an information rich measurement would be extremely sensitive to implementation errors.

From the general FI expression in Equation 5.12, we write the Fisher information for a mode-sorting system with a Gaussian aperture and the HG modes measurements as :

$$\mathcal{I}_{\mathrm{HG},q}(\theta) = \frac{N}{\sigma^2 q!} \left(q - \frac{\theta^2}{4\sigma^2} \right)^2 \left(\frac{\theta^2}{4\sigma^2} \right)^{q-1} \exp\left(-\frac{\theta^2}{4\sigma^2} \right), \tag{5.14}$$

Similarly, we write the Fisher information for a mode-sorting system with a hard aperture and the SB modes measurements as :

$$\mathcal{I}_{\mathrm{SB},q}(\theta) = \frac{4\pi^2 N}{\sigma^2}(1+2q) \left(\frac{q\sigma}{\pi\theta} j_q\left(\frac{\pi\theta}{\sigma}\right) - j_{q+1}\left(\frac{\pi\theta}{\sigma}\right) \right)^2 . \tag{5.15}$$

In a similar fashion to our analysis of the energy distribution across the modes, we now analyze the distribution of information across all the spatial modes. It is particularly interesting to consider the behavior of the information content for each mode in the limit of very small separation angles : $\theta \to 0^+$. Both the HG and SB mode basis follow similar trends :

$$\forall q \neq 1,\ \mathcal{I}_{\mathrm{HG},q}(\theta) \underset{\theta\to 0^+}{\rightarrow} 0, \qquad \text{and : } \mathcal{I}_{\mathrm{HG},1}(\theta) \underset{\theta\to 0^+}{\rightarrow} \frac{N}{\sigma^2} \tag{5.16}$$

$$\mathcal{I}_{\mathrm{SB},q}(\theta) \underset{\theta\to 0^+}{\rightarrow} 0, \qquad \text{and : } \mathcal{I}_{\mathrm{SB},1}(\theta) \underset{\theta\to 0^+}{\rightarrow} \frac{4\pi^2 N}{3\sigma^2} \tag{5.17}$$

We observe that only the 1^{st} mode captures all the available information at zero separation ($\theta = 0$) and remains the largest contributor in the sub-Rayleigh range ($\theta < \lambda/D$) as shown in Figure 5.11. It is worth highlighting that the information rich first mode effectively implements a differentiation between the two ASF superpositions on the left and the right side of the optical axis. Such a differential measurement provide optimal sensitivity in the $\theta \to 0$ regime as quantified in Equation 5.13. In contrast, the 0^{th} mode implements an average or summation of the left and right portions. It is also interesting to note the duality between the 0^{th} modes which capture the majority of the energy in the incident optical field while the 1^{st} modes capture the majority of the information in the narrow separations.

Beyond the 0^{th} and 1^{st} modes, we can extend the information analysis to the measurements of the Q first modes $\{m_0(\theta), \ldots, m_Q(\theta)\}$ using the information expression in Equation 5.12 as :

$$\mathcal{I}_{0,\ldots,Q}(\theta) = \sum_{q=0}^{Q} \mathcal{I}_q(\theta) = \sum_{q=0}^{Q} N \frac{m_q'(\theta)^2}{m_q(\theta)}, \tag{5.18}$$

we observe that the total information captured increases monotonically with the number of modes Q as shown in Figure 5.11.

Tsang *et al.* have shown that for the Gaussian ASF, measuring all of the HG modes yield a constant information value with respect to the separation (θ) which is linearly dependent to the total average number of photons (N) :

$$\forall \theta \in \mathbb{R}^+, \quad \sum_{q=0}^{+\infty} \mathcal{I}_{\mathrm{HG},q}(\theta) = \frac{N}{\sigma^2} \tag{5.19}$$

Similarly, we have shown that a similar result exists for the Sinc-Bessel modes (see Lemma 1 and Proof 3 in the Appendix). For a finite interval U of $\mathbb{R}^+$, we can write :

$$\forall \theta \in U, \quad \sum_{q=0}^{+\infty} \mathcal{I}_{\mathrm{SB},q}(\theta) = \frac{4N\pi^2}{3\sigma^2} \tag{5.20}$$

These information sums mirror the conservation of energy property in Equation 5.11 but for the information content. This is in stark contrast to the diminishing Fisher information for $\theta \to 0^+$ of a direct conventional imager employing an ideal detector array. Also, it should be noticed that for both measurement basis, collecting information at larger separation angles θ requires the collection of more higher order modes, as evident in Figure 5.11. This is expected as higher q-modes in the spatial HG and SB basis extend further away from the optical axis.

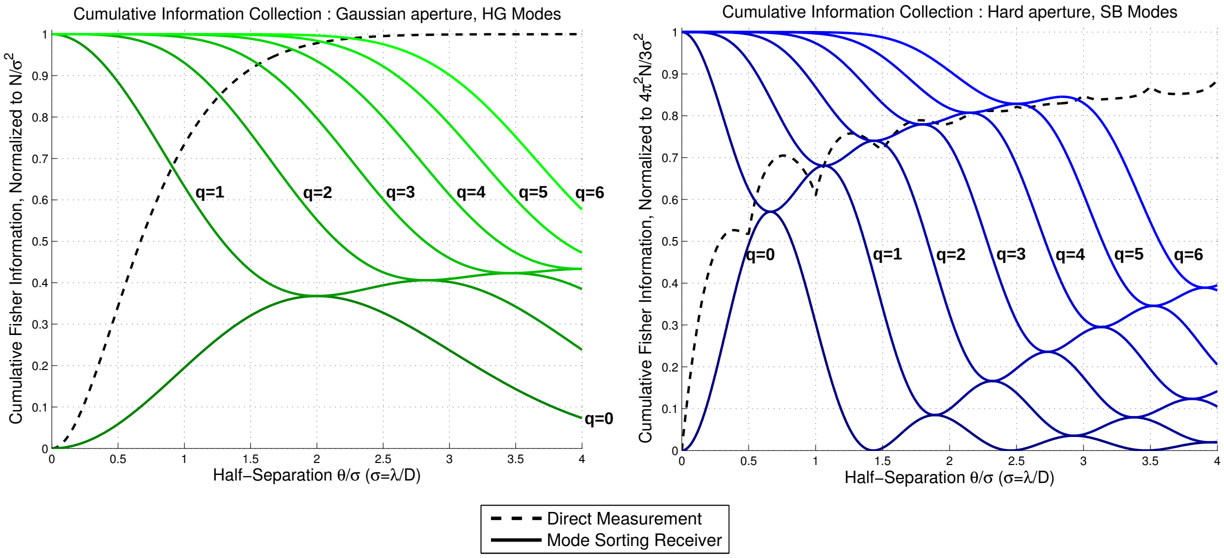


Figure 5.11: Fraction of the Fisher information collected over the cumulative measurements $0, \ldots, q$. **Left** : for a Gaussian aperture and HG modes. **Right** : for a hard aperture and SB modes. Similarly to the energy, the lowest order modes allow to acquire most of the information available at small separation angles θ.

5.3.3 BinSPADE Mode Measurement Design And Analysis

The alternate imaging architecture employing linear mode projections requires all the countably infinite modes comprising the mode basis to achieve the optimal constant performance over the entire range of angular separations. However, in the sub-Rayleigh range that is of practical interest here, because the conventional imager is optimal for larger separations, one only needs a relatively few mode measurements to overcome the performance limitation of the conventional imager and achieve optimal performance.

With this goal, Tsang *et al.* have proposed a straightforward Binary SPAtial-mode DEmultiplexing (BinSPADE) architecture [83] which measures any given single mode and its complement. Thus the first measurement $m_f(\theta)$ is the projection onto a single spatial mode $f(x)$ and the second is its residual $m_{r,f}(\theta) = 1 - m_f(\theta)$.

A good choice for the spatial mode measurement is dictated by the energy and information duality of the basis. In other words, one must choose a spatial mode which is efficient in energy collection or information collection. Thus we may choose either the 0^{th} mode, *i.e.* $f = h_0$ or s_0, which collects the energy and its residual collects the information, or the 1^{st} mode, *i.e.* $f = h_1$ or s_1, which collects the information and its residual collects the energy. We will refer to these two measurement designs as 0-BinSPADE when using the 0^{th} mode, or 1-BinSPADE when using the 1^{st} mode.

The Fisher information for a 0/1-BinSPADE can again be expressed analytically as the sum of the respective information for the mode and its residual. For a Gaussian aperture and HG mode basis we write the following expressions respectively for the 0 and 1-BinSPADEs :

$$\mathcal{I}_{\text{HG,0-BinSPADE}}(\theta) = \frac{N}{\sigma^2} \frac{\frac{\theta^2}{4\sigma^2}}{\exp\left(\frac{\theta^2}{4\sigma^2}\right) - 1}, \tag{5.21}$$

$$\mathcal{I}_{\text{HG,1-BinSPADE}}(\theta) = \frac{N}{\sigma^2} \frac{\left(1 - \frac{\theta^2}{4\sigma^2}\right)^2}{\exp\left(\frac{\theta^2}{4\sigma^2}\right) - \frac{\theta^2}{4\sigma^2}}. \tag{5.22}$$

Similarly, for a hard aperture and SB mode basis, we get :

$$\mathcal{I}_{\text{SB,0-BinSPADE}}(\theta) = \frac{4N}{\sigma^2} \frac{\left(\frac{\sigma}{\theta}\cos\left(\frac{\pi\theta}{\sigma}\right) - \frac{\sigma^2}{\pi\theta^2}\sin\left(\frac{\pi\theta}{\sigma}\right)\right)^2}{1 - j_0\left(\frac{\pi\theta}{\sigma}\right)^2}, \tag{5.23}$$

$$\mathcal{I}_{\text{SB,1-BinSPADE}}(\theta) = \frac{12N}{\sigma^2}\frac{\sigma^2}{\theta^2} \frac{\left[\sin\left(\frac{\pi\theta}{\sigma}\right) - 2j_1\left(\frac{\pi\theta}{\sigma}\right)\right]^2}{1 - 3j_1\left(\frac{\pi\theta}{\sigma}\right)^2}. \tag{5.24}$$

As we can observe from the plot of Fisher information for these two BinSPADE measurement designs in Figure 5.12, they will have a higher information content than a conventional imager for small angular separations (*i.e.* the sub-Rayleigh range $\theta \ll \sigma$). The location of the performance crossover between BinSPADEs and conventional imager occurs at separation angles equal to 250% and 170% of $\sigma = \lambda/D$ respectively for the 0 and 1-BinSPADEs of the HG mode basis and 120% and 80% of $\sigma = \lambda/D$ respectively for the 0 and 1-BinSPADEs of the SB mode basis. For larger separation angles, the information capacity of these architectures degrades to zero, as expected. It must also be noted that 0-BinSPADE significantly outperforms 1-BinSPADE for both Gaussian and hard aperture cases.

However, in a physical implementation of a BinSPADE measurement it is to be expected that non-idealities of physical components and/or imprecise opto-mechanics will degrade the performance from the performance predicted for an ideal BinSPADE measurement. Especially, given the energy-information duality of the 0^{th} and 1^{st} modes, one can expect to see a leakage from the direct measurement of a given mode over its residual because of an imperfect photon sorting implementation. This leakage effect can be modeled as the first measurement function $m_d(\theta)$ (either m_0 or m_1) being scaled by $(1-\epsilon)$ as it loses a small fraction $\epsilon > 0$ of the flux to its residual measurement (either $m_{r,0}$ or $m_{r,1}$) : $m_r(\theta) = 1 - (1-\epsilon)m_d(\theta)$. For simplicity, we define the efficiency $\rho = 1 - \epsilon$ and derive the following expressions for the Fisher information of both 0- and 1-BinSPADE conditioned on this leakage parameter. For

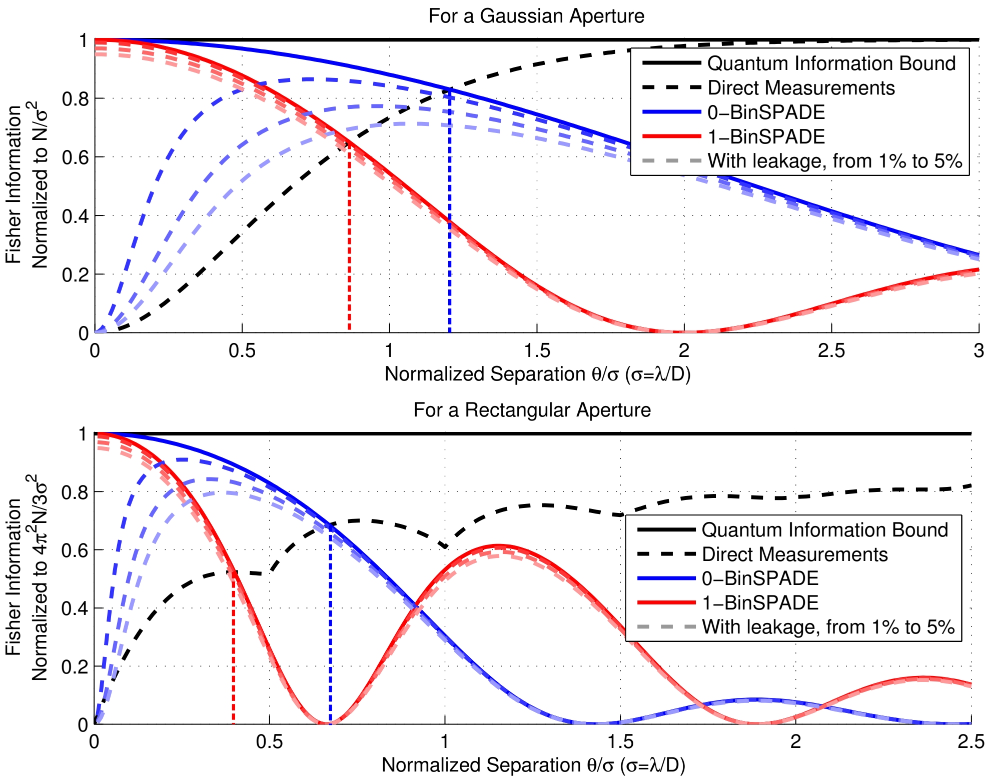


Figure 5.12: Performance comparison of Direct Measurements versus 0 and 1-BinSPADE receivers for Gaussian (**top**) and hard (**bottom**) apertures. The y-axes are normalized to the respective Quantum Fisher information bounds.

a Gaussian aperture we write :

$$\mathcal{I}_{\text{HG,0-BinSPADE},\rho}(\theta) = \frac{N\rho}{\sigma^2}\frac{\frac{\theta^2}{4\sigma^2}}{\exp\left(\frac{\theta^2}{4\sigma^2}\right)-\rho}, \tag{5.25}$$

$$\mathcal{I}_{\text{HG,1-BinSPADE},\rho}(\theta) = \frac{N\rho}{\sigma^2}\frac{\left[1-\frac{\theta^2}{4\sigma^2}\right]^2}{\exp\left(\frac{\theta^2}{4\sigma^2}\right)-\rho\frac{\theta^2}{4\sigma^2}}. \tag{5.26}$$

Similarly, for a hard aperture and SB modes, we write :

$$\mathcal{I}_{\text{SB,0-BinSPADE},\rho}(\theta) = \frac{4N\rho}{\sigma^2}\frac{\left(\frac{\sigma}{\theta}\cos\left(\frac{\pi\theta}{\sigma}\right)-\frac{\sigma^2}{\pi\theta^2}\sin\left(\frac{\pi\theta}{\sigma}\right)\right)^2}{1-\rho j_0\left(\frac{\pi\theta}{\sigma}\right)^2}, \tag{5.27}$$

$$\mathcal{I}_{\text{SB,1-BinSPADE},\rho}(\theta) = \frac{12N\rho}{\sigma^2}\frac{\sigma^2}{\theta^2}\frac{\left[\sin\left(\frac{\pi\theta}{\sigma}\right)-2j_1\left(\frac{\pi\theta}{\sigma}\right)\right]^2}{1-3\rho j_1\left(\frac{\pi\theta}{\sigma}\right)^2}. \tag{5.28}$$

Despite the superior information performance of the ideal 0-BinSPADE over the 1-BinSPADE, we find that the information of the former collapses immediately to zero at $\theta = 0$ and for any leakage $\epsilon > 0$ (any efficiency $\rho < 1$). On the contrary, the latter only degrades linearly with the efficiency ρ and thus more resilient to this particular defect.

5.4 Mean Square Error Analysis of Mode Measurements

While a Fisher information analysis of mode measurements provide a lower bound on the error in estimating the separation angle via the Cramer-Rao bound, the actual performance of an estimator is equally important to analyze. For such an analysis , we have to choose an estimator of the separation θ from a set of mode measurements. Naturally, the maximum likelihood estimator (MLE) is a good choice given that it is an asymptotically efficient. The generalized definition of the MLE in the case of $Q+1$ statistically independent measurements $\{y_0, \ldots, y_Q\}$ can be expressed as :

$$\widehat{\theta}_{\text{MLE}}(y_0, \ldots, y_Q) = \arg\max_{\theta} \left\{ \prod_{q=0}^{Q} \mathcal{P}(y_q | N m_q(\theta)) \right\}. \tag{5.29}$$

This expression is based on the fact that each modal measurement Y_q follows a Poisson distribution, *i.e.* $Y_q \sim \mathcal{P}\left(N m_q(\theta)\right)$ assuming that the total average photon flux N is known. The MLE can be expressed in terms of the log-likelihood for a single mode measurement as :

$$\begin{aligned} \widehat{\theta}_{\text{MLE}}(y_q) &= \arg\max_{\theta} \left\{ y_q \ln(N m_q(\theta)) - N m_q(\theta) \right\} \\ &= m_q^{-1} \circ R_U \left(\frac{y_q}{N} \right), \end{aligned} \tag{5.30}$$

where m_q^{-1} denotes the inverse of m_q, which exists over some fixed interval of interest U containing $\theta = 0$. R_U is a simple restriction operator to the image V of U by m_q, it clamps overflowing values of the quotient y_q/N to the closest extrema of V in order to avoid indefinite portions of m_q^{-1}. And $\circ$ denotes the composition operator, *i.e.* $m_q^{-1} \circ R_U(\cdot) = m_q^{-1}(R_U(\cdot))$.

Additionally, we can also express the estimator analytically in case of a BinSPADE architecture : we note y_q and $y_{r,q}$ the respective outputs of $m_q(\theta)$ and its residual $1 - m_q(\theta)$, we write :

$$\begin{aligned}\widehat{\theta}_{\text{MLE}}(y_q, y_{r,q}) &= \arg\max_{\theta} \left\{\mathcal{P}\left(y_q \middle| Nm_q(\theta)\right) \mathcal{P}\left(y_{r,q} \middle| N - Nm_q(\theta)\right)\right\} \\ &= m_q^{-1} \circ R_U\left(\frac{y_q}{y_q + y_{r,q}}\right),\end{aligned} \tag{5.31}$$

Note that this estimator is similar to the one for a single Q-mode (Equation 5.30) in the limit of large N where $y_q + y_{r,q}$ becomes a good approximation of the total flux (*i.e.* $y_q + y_{r,q} \sim \mathcal{P}(N)$).

Given the ML estimator definition, we can now compute the normalized RMSE (NRMSE) performance of a BinSPADE measurement via Monte Carlo simulations. For this measurement architecture we have noted that only small angular separations (below Rayleigh limit) are of interest when compared to a direct imager, we thus restrict θ to a range extending only slightly beyond Rayleigh's limit ($\sigma/2$) and where the measurement function is invertible. We repeat the NRMSE performance analysis for several values of the total flux N. Ranging from 10 to $10,000$ photons. Note that while the Fisher information only scales linearly with the parameter (θ), we expect a more complex behavior of the ML estimator which becomes strongly biased in the small photon number regime.

The MLE performance results are summarized in Figures 5.13 and 5.14 for 0 and 1-BinSPADE measurements with a Gaussian aperture and a hard aperture respectively. In each figure we include the NCRLB to provide a reference for the evaluation of NRMSE performance. We also include the $\text{NRMSE}_{\text{Direct}}$ and $\text{NCRLB}_{\text{Direct}}$ of the direct imager employing an ideal continuum FPA. One can immediately note that in a high flux regime (high N), the performance of the BinSPADE measurement is nearly identical to the Cramer-Rao

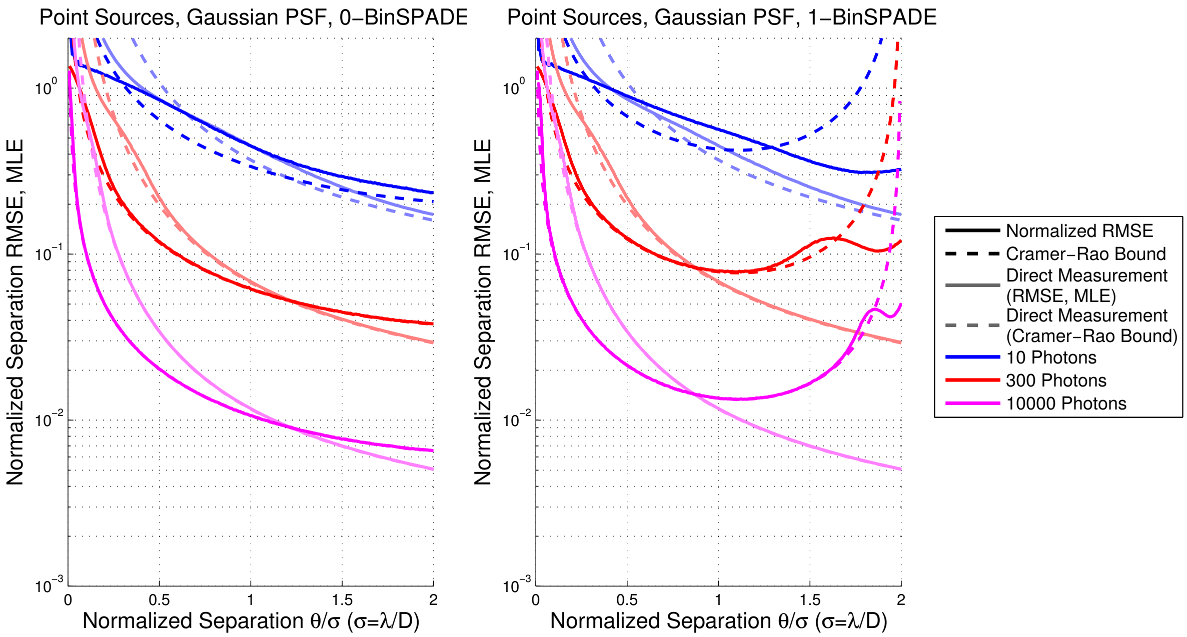


Figure 5.13: Normalized RMSE plots from Monte Carlo-Estimation estimated performance of MLE. **Left** : 0-BinSPADE for Gaussian aperture. **Right** : 1-BinSPADE for Gaussian aperture. The estimators for both can be written analytically (with Lambert-W function for the 1-BinSPADE) over the range of θ shown, where the associated measurement function is invertible.

lower bound. On the other hand, the estimator performance deviates from the NCRLB in the photon-starved regime (low N). In this case, it is interesting to note that NRMSE can actually become better than the NCRLB due to bias. This is particularly visible for 1-BinSPADEs in the large angular separation (θ) limit. However, it is important to note that the MLE enters this super-efficient regime because we constrain its output value to be within the fixed range U as written in Equation 5.31.

Furthermore, one can observe that the BinSPADEs measurements can significantly outperform the conventional direct imager design. To quantify the improvement with reference to the the direct imager we can define the gain G as :

$$G = 20\log_{10}(\mathrm{NCRLB}_{\mathrm{Direct}}\,/\,\mathrm{NRMSE}), \tag{5.32}$$

where positive values indicate superior performance. At $N = 10,000$ photons, for the Gaussian aperture and both 0,1-BinSPADEs achieve a performance gain of approximately

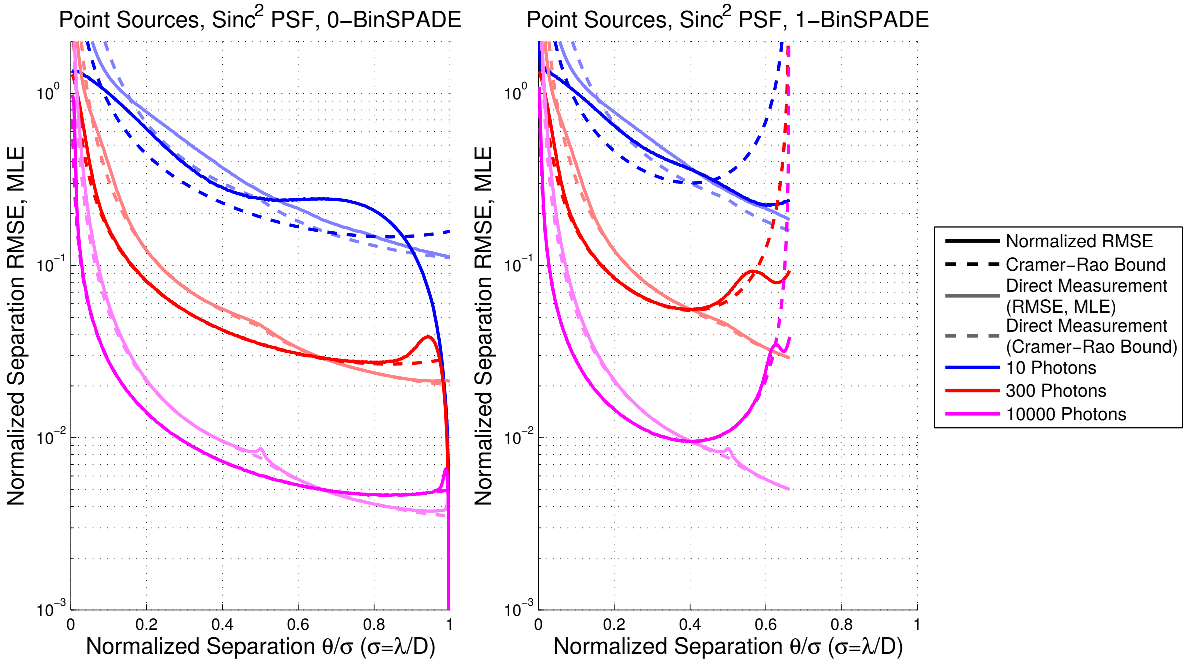


Figure 5.14: Normalized RMSE plots from Monte-Carlo Estimation. **Left** : 0-BinSPADE for hard aperture. **Right** : 1-BinSPADE for hard aperture. Note that the θ/σ range restriction for 1-BinSPADE corresponds to the presence of the first zero in the Fisher information.

+4dB at $\theta = 0.2\sigma$ and +8dB at $\theta = 0.1\sigma$ respectively. Similarly, for the hard aperture we observe a gain of nearly +3dB at $\theta = 0.2\sigma$ and +5dB at $\theta = 0.1\sigma$. Finally, it is important to point out that the BinSPADEs measurement, while gathering finite information near $\theta = 0$, do not achieve a constant relative accuracy for all small angles as evident from the NRMSE plot, but achieve a performance that scales as $\mathcal{O}(1/\theta\sqrt{N})$. In comparison, the direct imager with a Gaussian aperture achieves a performance scaling of $\mathcal{O}(1/\theta^2\sqrt{N})$ and $\mathcal{O}(1/\theta^{\frac{3}{2}}\sqrt{N})$ in the case of a hard aperture, which are fundamentally inferior to the BinSPADE.

5.5 Efficient Binary SPADE For Arbitrary ASF

In the previous sections, we have limited our discussion and analysis to Gaussian and hard apertures, that produce Gaussian and cardinal sine ASFs respectively. We have already commented on the common themes among performance of these two apertures for direct imager and BinSPADE measurements. Nonetheless, we have noted that the system performance does depend on the actual pupil function and the BinSPADE design needs to be adapted to the optimal choice of mode basis given an aperture function. To better understand how the aperture shape impacts the characteristics and performance of the BinSPADEs, we need to generalize this measurement scheme to arbitrary ASF and the associated optimal mode basis.

We begin by noting the fact that, for both of the aperture functions considered earlier, the 0th spatial mode matches the ASF. From this observation, we can develop a general measurement expression for a properly normalized complex-valued ASF $A_\sigma(x)$ in terms of its autocorrelation function Γ_A defined as :

$$\Gamma_A(x') = \int_{-\infty}^{+\infty} \overline{A(x)} A(x+x')\; dx. \tag{5.33}$$

Thus we can modify the measurement function in Equation 5.6 where the spatial mode is the ASF itself :

$$m_{A,0}(\theta) = \left|\Gamma_A\left(\frac{\theta}{\sigma}\right)\right|^2. \tag{5.34}$$

One can particularly verify that in the limit $\theta \to 0$ we have $m_{A,0}(\theta) = 1$, *i.e.* this ASF mode measurement collects all of the available energy at $\theta = 0$. The corresponding generalized residual measurement function takes the obvious form : $m_{A,r,0}(\theta) = 1 - m_{A,0}(\theta)$. Similarly to previous BinSPADE measurement designs, two shot-noise limited detectors are placed at the output of each mode projection and measure Poisson-distributed data with mean values $Nm_{A,0}(\theta)$ and $Nm_{A,r,0}(\theta)$ respectively. From Equation 5.12, we can infer that the Fisher information of this generalized 0-BinSPADE receiver is :

$$\mathcal{I}_{\text{Gen,0-BinSPADE}}(\theta) = \frac{4N}{\sigma^2} \frac{\text{Re}\left[\overline{\Gamma_A^{(1)}\left(\frac{\theta}{\sigma}\right)}\Gamma_A\left(\frac{\theta}{\sigma}\right)\right]^2}{\left|\Gamma_A\left(\frac{\theta}{\sigma}\right)\right|^2\left(1-\left|\Gamma_A\left(\frac{\theta}{\sigma}\right)\right|^2\right)}, \tag{5.35}$$

where $\Gamma_A^{(1)}$ is the first derivative of Γ_A. If we now assume that the autocorrelation Γ_A admits the following second order expansion around $\theta = 0$, constrained by its Hermitian symmetry and maximum property :

$$\Gamma_A(x) \underset{\theta \to 0}{=} 1 + i\beta x - \frac{\alpha}{2}x^2 + \mathcal{O}(x^3), \tag{5.36}$$

with $\alpha \geq 0$ and $\beta \in \mathbb{R}$, then we can derive the limiting value of the Fisher information expression in Equation 5.35 in the narrow angular separation regime ($\theta \to 0$) :

$$\mathcal{I}_{\text{Gen,0-BinSPADE}}(\theta) \underset{\theta \to 0}{\to} \frac{4N}{\sigma^2}(\alpha - \beta^2). \tag{5.37}$$

One can immediately note that this limiting value is again linearly proportional to the average number of photons N and that for purely real ASFs ($\beta = 0$) the limit is improved. It is also reassuring to observe that Equation 5.37 reduces to known values for the Gaussian ASF ($\alpha = 1/4$, $\beta = 0$, $\mathcal{I} = N/\sigma^2$) and the cardinal sine ASF ($\alpha = \pi^2/3$, $\beta = 0$, $\mathcal{I} = 4\pi^2 N/3\sigma^2$).

More interestingly, note that an ASF with high spatial variability will have its autocorrelation function $\Gamma_A(x)$ quickly decreasing to zero for small shift x. This indicates that the autocorrelation function has a sharp curvature in $\theta = 0$, and thus a large α coefficient in the expansion of Equation 5.36. Consequently, a mode-sorting measurement device with an ASF having pronounced spatial features is more sensitive to small shifts of the point-sources ASFs and therefore, it collects a large amount of information about the angular separation as a result of Equation 5.37. For example, the cardinal sine ASF has more spatial features than the smooth Gaussian ASF and this leads to the difference in Fisher information content between Gaussian and hard apertures written in Equations 5.19 and 5.20 and also shown for conventional direct imager in Figure 5.5.

A second property of the mode-sorting measurement device can be obtained from the properties of the autocorrelation. Namely, the Wiener–Khinchin theorem states that the Fourier Transform of the autocorrelation Γ_A is equal to the absolute square of the Fourier Transform of $A(x)$, *i.e.* $\mathcal{F}\left[\Gamma_A(x)\right] = \left|\mathcal{F}\left[A(x)\right]\right|^2$. Hence, as the ASF is related to the Fourier transform of the aperture function the ASF autocorrelation function is equal to the absolute square of the (spatially inverted) aperture function : *i.e.* $\Gamma_A(x) = |T(-x)|^2$. Therefore this system is insensitive to the phase component of the pupil function and thus, to any phase-induced optical aberrations known in the system (for example : defocusing, spherical aberration, coma, *e.t.c.*). This is in contrast to the traditional direct imager for which the Fisher information is decreasing in the presence of optical aberrations. More generally, we know that the optical field in the image plane is related to the optical field in the pupil plane via the Fourier transform, which is a change of basis in the $\mathcal{L}_2$ vector space. As the spatial mode demultiplexer is only performing linear projections optically in the image plane, it can be adapted to work directly in the pupil plane by inverting this linear field propagation transform. For instance in the case of a hard aperture, the system is performing projections against the Sinc-Bessel spatial modes in the image plane which is equivalent to performing projections against the Legendre spatial modes directly in the pupil plane as the two basis are Fourier conjugates. With this modification, we can emphasize that the mode measurements can be performed directly from the entrance pupil of the system objective, which is thus irrelevant for this imaging task. Therefore, we should expect that the inherent angular resolution ($\approx \lambda/D$) and optical aberrations are also irrelevant to the measurements performance.

After having described the 0th generalized mode measurement in Equation 5.34, it is now possible to construct an orthonormal basis to perform generalized parallel mode decomposition by using the derivatives of the ASF $A(x)$, assuming it is infinitely differentiable. To develop such a mode basis, it is useful to consider the following identity for successive derivatives of the autocorrelation function :

$$\int_{-\infty}^{+\infty} \overline{A^{(q)}(x)} A(x+x')\, dx = (-1)^q \Gamma_A^{(q)}(x') \tag{5.38}$$

where $A^{(q)}(x)$ is the q^{th} order derivative of $A(x)$. While the set of the ASF and its successive derivatives is not necessarily orthogonal, we can use Gram-Schmidt orthonormalization to select complex weights $\omega_{k,q}$ and to construct an orthonormal set of modes $\{M_q(x)\}$

which can be expressed, with the explicit spatial normalization of scale σ, as :

$$\frac{1}{\sqrt{\sigma}} M_q\left(\frac{x}{\sigma}\right) = \sum_{k=0}^{q} (-1)^k \frac{\omega_{k,q}}{\sigma^{q+\frac{1}{2}}} A^{(q)}\left(\frac{x}{\sigma}\right). \tag{5.39}$$

Note that we may have to remove any zero-functions in this set as necessary. We also always have $\omega_{0,0} = 1$ because of the normalization of $A(x)$. Given this mode basis design, we find back the 0th generalized mode, *i.e.* the ASF, and subsequently develop the 1st generalized mode with the appropriate spatial normalization :

$$\frac{1}{\sqrt{\sigma}} M_0\left(\frac{x}{\sigma}\right) = \frac{1}{\sqrt{\sigma}} A\left(\frac{x}{\sigma}\right) \tag{5.40}$$

$$\frac{1}{\sqrt{\sigma}} M_1\left(\frac{x}{\sigma}\right) = \frac{1}{\sqrt{\sigma}} \frac{-A^{(1)}\left(\frac{x}{\sigma}\right) + \Gamma_A^{(1)}(0) A\left(\frac{x}{\sigma}\right)}{\sqrt{-\Gamma_A^{(2)}(0) - \left|\Gamma_A^{(1)}(0)\right|^2}}, \tag{5.41}$$

where $A^{(1)}(x)$ is the first derivative of the ASF, $\Gamma_A^{(1)}$ and $\Gamma_A^{(2)}$ are respectively the first and second derivatives of the autocorrelation function. As the autocorrelation is Hermitian, it implies that $\Gamma_A^{(1)}(0)$ is purely imaginary. Thus in the case of a purely real ASF this term is equal to zero and the first mode can be constructed with only the first derivative of the ASF, as it is indeed the case for both the Gaussian and cardinal sine examples.

With these modes and the property given in Equation 5.38, the mode projection function of Equation 5.6 can be generalized to :

$$m_{A,q}(\theta) = \frac{1}{2}\left|\sum_{k=0}^{q} \frac{\overline{\omega_{k,q}}}{\sigma^k} \Gamma_A^{(k)}\left(\frac{\theta}{\sigma}\right)\right|^2 + \frac{1}{2}\left|\sum_{k=0}^{q} \frac{\overline{\omega_{k,q}}}{\sigma^k} \Gamma_A^{(k)}\left(-\frac{\theta}{\sigma}\right)\right|^2. \tag{5.42}$$

The Fisher information of such mode measurement can be computed using Equation 5.12. Of course, for the 0th mode we arrive at the same expression as in Equation 5.34 for the measurement and Equation 5.35 for the 0-BinSPADE information. More interestingly, for the generalized 1st mode we obtain :

$$m_{A,1}(\theta) = \frac{\left|\Gamma_A^{(1)}\left(\frac{\theta}{\sigma}\right) - \Gamma_A^{(1)}(0)\Gamma_A\left(\frac{\theta}{\sigma}\right)\right|^2}{-\Gamma_A^{(2)}(0) - \left|\Gamma_A^{(1)}(0)\right|^2}, \tag{5.43}$$

and :

$$\mathcal{I}_{\text{Gen,1-BinSPADE}}(\theta) = \frac{4N}{\sigma^2} \frac{\text{Re}\left[\overline{f^{(1)}\left(\frac{\theta}{\sigma}\right)} f\left(\frac{\theta}{\sigma}\right)\right]^2}{\left|f\left(\frac{\theta}{\sigma}\right)\right|^2 \left(-f^{(1)}(0) - \left|f\left(\frac{\theta}{\sigma}\right)\right|^2\right)}, \tag{5.44}$$

$$\text{with: } f\left(\frac{\theta}{\sigma}\right) = \Gamma_A^{(1)}\left(\frac{\theta}{\sigma}\right) - \Gamma_A^{(1)}(0)\Gamma_A\left(\frac{\theta}{\sigma}\right). \tag{5.45}$$

Given that the autocorrelation admits an expansion as Equation 5.36, we can show that the Fisher information of this 1-BinSPADE also tends toward the same limit value achieved by the 0-BinSPADE in Equation 5.37 :

$$\mathcal{I}_{\text{Gen,1-BinSPADE}}(\theta) \underset{\theta \to 0}{\to} \frac{4N}{\sigma^2}(\alpha - \beta^2). \tag{5.46}$$

Finally, to complete the analysis of imperfect BinSPADE measurements we can introduce the previously described leakage term $\epsilon > 0$, or equivalently the efficiency $\rho < 1$, to test the resilience of both generalized architectures. By transforming the modes output to take into account this imperfection, we obtain for the Fisher information of the generalized leaky 0-BinSPADE and 1-BinSPADE respectively :

$$\mathcal{I}_{\text{Gen,0-BinSPADE},\rho}(\theta) = \frac{4N\rho}{\sigma^2} \frac{\text{Re}\left[\overline{\Gamma_A^{(1)}\left(\frac{\theta}{\sigma}\right)} \Gamma_A\left(\frac{\theta}{\sigma}\right)\right]^2}{\left|\Gamma_A\left(\frac{\theta}{\sigma}\right)\right|^2 \left(1 - \rho\left|\Gamma_A\left(\frac{\theta}{\sigma}\right)\right|^2\right)}, \tag{5.47}$$

$$\mathcal{I}_{\text{Gen,1-BinSPADE},\rho}(\theta) = \frac{4N\rho}{\sigma^2} \frac{\text{Re}\left[\overline{f^{(1)}\left(\frac{\theta}{\sigma}\right)} f\left(\frac{\theta}{\sigma}\right)\right]^2}{\left|f\left(\frac{\theta}{\sigma}\right)\right|^2 \left(-f^{(1)}(0) - \rho\left|f\left(\frac{\theta}{\sigma}\right)\right|^2\right)}. \tag{5.48}$$

From the properties of the autocorrelation (particularly that $\Gamma_A(0) = 1$ and $\Gamma_A^{(1)}(0)$ is purely imaginary), we can infer that the numerator of Equation 5.47 always tends to zero as $\theta \to 0$, but that the denominator only does so if the efficiency ρ is exactly equal to unity. Thus, for any sub-unity efficiency implementation ($\rho < 1$), the 0-BinSPADE is not able to capture any information for small separation angle : $\mathcal{I}_{\text{0-BinSPADE},\rho<1}(\theta) \to 0$ near $\theta = 0$. On the other hand, one can develop the limit value of FI in Equation 5.48 as :

$$\mathcal{I}_{\text{Gen,1-BinSPADE},\rho}(\theta) \underset{\theta\to 0}{\to} \frac{4N\rho}{\sigma^2}(\alpha - \beta^2). \tag{5.49}$$

Thus, the generalized 1-BinSPADE is resilient to this particular leakage modeled in the deep sub-Rayleigh limit, as its performance only decrease linearly with the efficiency (ρ).

5.6 Candidate Architectures : Performance Analysis

In this section we consider two candidate mode-decomposition implementations and analyze their performance in terms of the two point-sources separation problem. The first architecture describes an interferometric setup proposed by Tsang *et al.* and which has been implemented by several research teams. For this architecture, we analyze its performance for an arbitrary ASF and identify its inherent strengths and limitations. The second architecture implements a single-mode measurement along with a flux measurement using a beam splitter and a volume hologram. We analyze its binary mode design and performance as well.

5.6.1 Architecture : Mach-Zehnder Interferometer

In [87], Tsang *et al.* proposed an interesting architecture that implements two-modes decomposition based on a Mach-Zehnder interferometer. This architecture has been since implemented and its performance quantified by Tang *et al.* [88] and Tham *et al.* [89], both for a Gaussian aperture. In this mode decomposition architecture, instead of decomposing the optical field in the image plane directly into a series of modes, the two incoherent electric fields are first split between the two balanced arms of the interferometer. In the first arm,

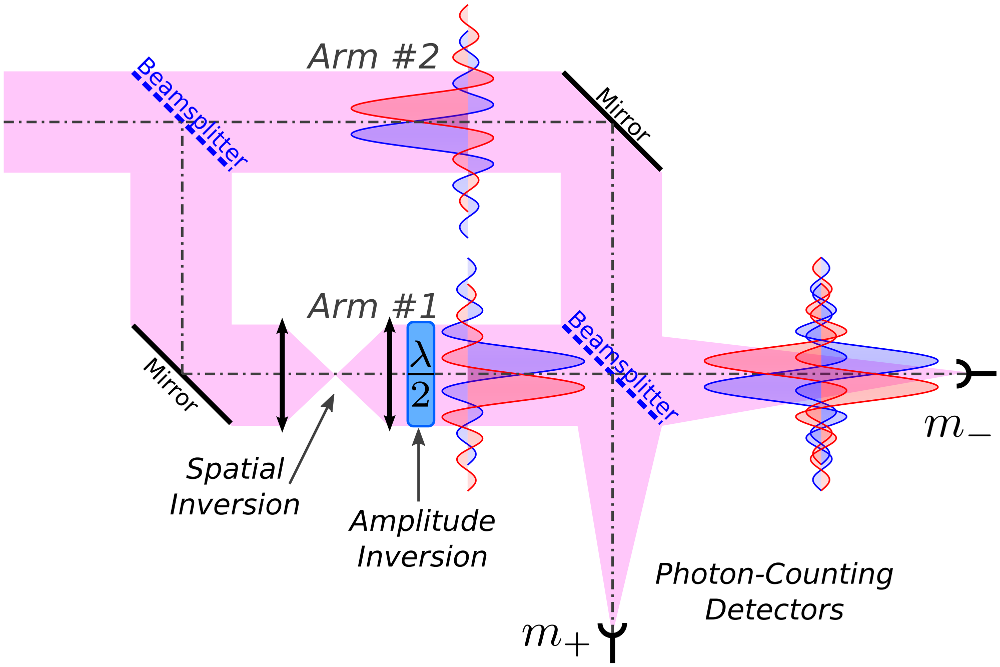


Figure 5.15: SLIVER/Interferometric implementation layout.

the optical field is first inverted spatially, using a $2f$ system, then a phase-retarder is used to invert its amplitude before being recombined with the components of the original field from the second arm, as shown in Figure 5.15.

Let us denote the two measurements as $m_+(\theta)$ and $m_-(\theta)$ at the two output ports of the interferometer. If we employ ideal 50/50 beam splitters, shot-noise limited photodetectors and ignore any phase mismatch or spatial distortions that may arise from a non-ideal component we obtain the following mathematical description of m_+ and m_- :

$$\begin{aligned} m_-(\theta) = &\underbrace{\int_{-\infty}^{+\infty} \left| \frac{iA_\sigma(x-\theta)}{2\sqrt{2}} - \frac{iA_\sigma(-x-\theta)}{2\sqrt{2}} \right|^2 dx}_{\text{Source 1}} \\ &+ \underbrace{\int_{-\infty}^{+\infty} \left| \frac{iA_\sigma(x+\theta)}{2\sqrt{2}} - \frac{iA_\sigma(-x+\theta)}{2\sqrt{2}} \right|^2 dx}_{\text{Source 2}} \\ = \frac{1}{2} - \frac{1}{4}\mathrm{Re}&\left[\int_{-\infty}^{+\infty} \overline{A_\sigma(x-\theta)} A_\sigma(-x-\theta) + \overline{A_\sigma(x+\theta)} A_\sigma(-x+\theta)\, dx \right]. \end{aligned} \tag{5.50}$$

In the case where the ASF is real and symmetric (*i.e.* $A(-x) = \overline{A(x)}$), one can verify that $m_{-}(0) = 0$. Similarly, we can write the second as :

$$
\begin{aligned}
m_{+}(\theta) = & \underbrace{\int_{-\infty}^{+\infty} \left| \frac{A_\sigma(x-\theta)}{2\sqrt{2}} \frac{A_\sigma(-x-\theta)}{2\sqrt{2}} \right|^2 \, dx}_{\text{Source 1}} \\
& + \underbrace{\int_{-\infty}^{+\infty} \left| \frac{A_\sigma(x+\theta)}{2\sqrt{2}} + \frac{A_\sigma(-x+\theta)}{2\sqrt{2}} \right|^2 \, dx}_{\text{Source 2}} \\
= & \frac{1}{2} + \frac{1}{4} \mathrm{Re} \left[\int_{-\infty}^{+\infty} \overline{A_\sigma(x-\theta)} A_\sigma(-x-\theta) + \overline{A_\sigma(x+\theta)} A_\sigma(-x+\theta) \right] \, dx. \qquad (5.51)
\end{aligned}
$$

Again, if the ASF is real and symmetric, one can verify that $m_{+}(0) = 1$. This approach hence offers an elegant self-referencing mechanism to generate measurements of the generalized 0^{th} mode but only if the ASF is real and symmetric. The two output are considered to be independent and Poisson-distributed with means $Nm_{-}(\theta)$ and $Nm_{+}(\theta)$ respectively, for a total average photons flux N. Their respective Fisher information can therefore be expressed as (similarly to Equation 5.12) :

$$
\begin{aligned}
\mathcal{I}_{-}(\theta) = & \frac{N}{16 m_{-}(\theta)} \mathrm{Re} \left[\int_{-\infty}^{+\infty} \overline{A'_\sigma(x-\theta)} A_\sigma(-x-\theta) + \overline{A_\sigma(x-\theta)} A'_\sigma(-x-\theta) \right. \\
& \left. - \overline{A'_\sigma(x+\theta)} A_\sigma(-x+\theta) - \overline{A_\sigma(x+\theta)} A'_\sigma(-x+\theta) \, dx \right]^2, \qquad (5.52) \\
\mathcal{I}_{+}(\theta) = & \frac{N}{16 m_{+}(\theta)} \mathrm{Re} \left[\int_{-\infty}^{+\infty} \overline{A'_\sigma(x-\theta)} A(-x-\theta) + \overline{A_\sigma(x-\theta)} A'_\sigma(-x-\theta) \right. \\
& \left. - \overline{A'_\sigma(x+\theta)} A_\sigma(-x+\theta) - \overline{A_\sigma(x+\theta)} A'_\sigma(-x+\theta) \, dx \right]^2. \qquad (5.53)
\end{aligned}
$$

The total information provided by the two measurements is the sum of each : $\mathcal{I}_{\text{MZ}}(\theta) = \mathcal{I}_{-}(\theta) + \mathcal{I}_{+}(\theta)$. For Gaussian and hard apertures, the FI performance is similar to that of a 1-BinSPADE, as shown in Figure 5.16, in that it degrades quickly for angles larger than Rayleigh's limit. Overall, this implementation is a suitable approach for two-modes measurement as long as the pupil function is purely real and symmetric (*i.e.* an even function) which yields an ASF that is real and symmetric as well. Any deviation from a symmetric

ASF will incur a performance penalty, especially near $\theta = 0$.

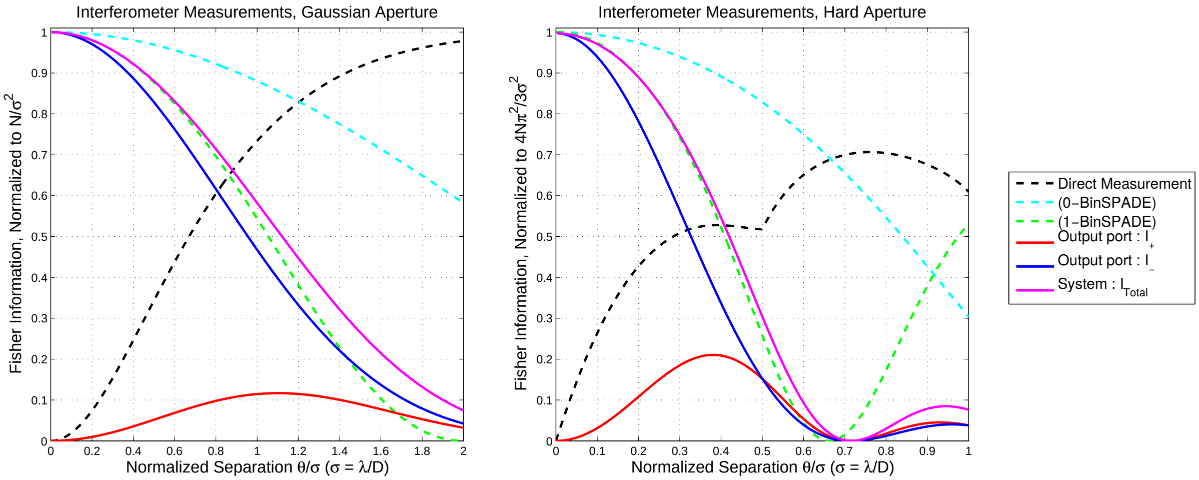


Figure 5.16: Fisher information for the interferometric measurements. The curves for the 0 and 1-BinSPADE offer a comparison to the two-outputs mode-sorting receiver.

5.6.2 Unimodal Architecture : Volume Hologram

In order to reduce the complexity of the BinSPADE mode decomposition measurement, we consider an implementation of a single-mode projection through a self-referencing mechanism such as a volume hologram (VH) as shown in Figure 5.16. The VH can be designed so as to direct certain components of the optical field in a given direction and with a very high transmission efficiency ($> 99\%$). Following our descriptions of the mode properties in the previous discussion, we can design the VH to either selectively propagate the 1$^{\text{st}}$ mode or the residual of the 0$^{\text{th}}$ mode as we have shown that they are the only two measurements that contribute to the majority of available information in the sub-Rayleigh regime. In this discussion we will choose the 1$^{\text{st}}$ mode ($m_1(\theta)$) and consider a hard aperture.

Beyond the first mode we also need to measure the average total flux N incident on the imager aperture to be able to estimate the actual separation via the MLE defined in Equation 5.30. To perform this flux estimation in parallel to the mode measurement, we distribute the total flux available between the two paths of a beam splitter, positioned near the exit pupil, as shown in Figure 5.17. This will of course degrade the system performance

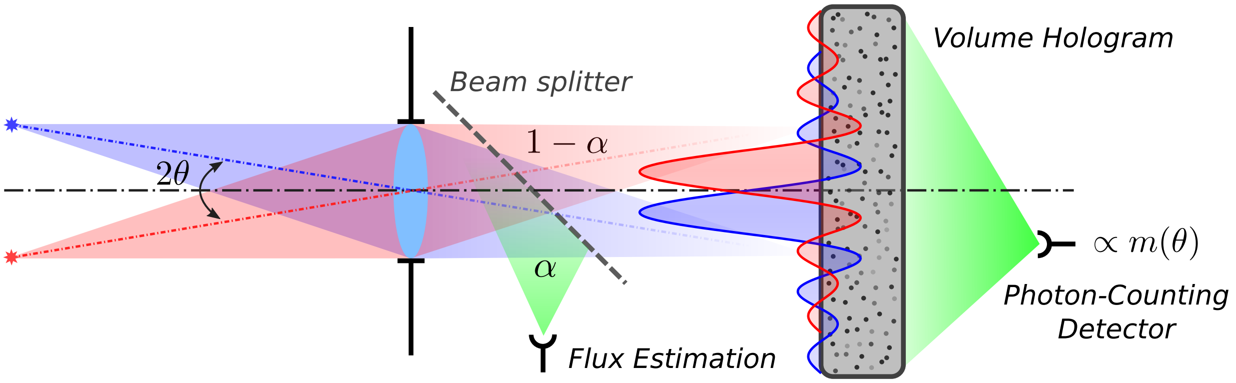


Figure 5.17: Proposed single-mode measurement architecture.

as some fraction of the total photons that would have contributed to the single mode measurement of interest are now diverted toward the measurement of total photon flux N. Thus, in order to optimize the information throughput of such a measurement design we must find the optimal beam splitter ratio α.

For this analysis, α denotes the fraction of the intensity allocated to the flux measurement, *i.e.* the beam splitter ratio is α:$1-\alpha$. The flux measurement y_N is thereby obtained at the Poisson-corrupted output of mean αN. The mode measurement y_M is independent and Poisson distributed of mean $(1-\alpha)Nm_1(\theta)$. Now we can define the MLE given the two measurements y_M and y_N as :

$$\widehat{\theta}_{\text{Single,MLE},\alpha}(y_N, y_M) = m_1^{-1} \circ R_U \left(\frac{\alpha y_M}{(1-\alpha)y_N} \right) \tag{5.54}$$

To optimize the beam splitter ratio α, we assume no prior knowledge of the angular separation and thus can seek an optimal α_{Opt} which minimizes the normalized error across a range $[0, \theta_{\max}]$, for a given average flux N, *i.e.* :

$$\alpha_{\text{Opt}}(N) = \underset{\alpha \in (0,1)}{\arg\min} \left\{ \int_0^{\theta_{\max}} \text{NRMSE}_{\text{single}}(\theta, \alpha, N)\ d\theta \right\}, \tag{5.55}$$

where $\text{NRMSE}_{\text{single}}(\theta, \alpha, N)$ denotes the NRMSE of the MLE in Equation 5.54. We pursue a numerical optimization and find that α_{Opt} is relatively insensitive to the incident flux N, as shown in Figure 5.18. Especially, we find that the optimal α is 0.29, 0.27 and 0.28 for $N = 50$, $N = 300$ and $N = 10,000$ photons respectively. We also compare the normalized estimate error performance for this design against an ideal design in which all the photons would be available for the single mode measurement and the flux N is known a priori. We observe that this single mode measurement maintains a significant performance improvement over the conventional direct imager for $\theta < 0.2\lambda/D$. For example at $N = 10,000$ photons, the relative error for the single mode measurement is half of the direct imager at $\theta \approx 0.1\lambda/D$; and about 35% smaller at $N = 50$ photons. At $N = 10,000$ photons, the relative error for the single mode measurement is almost a quarter of direct imager at $\theta \approx 0.05\lambda/D$; and about 30% smaller at $N = 50$ photons.

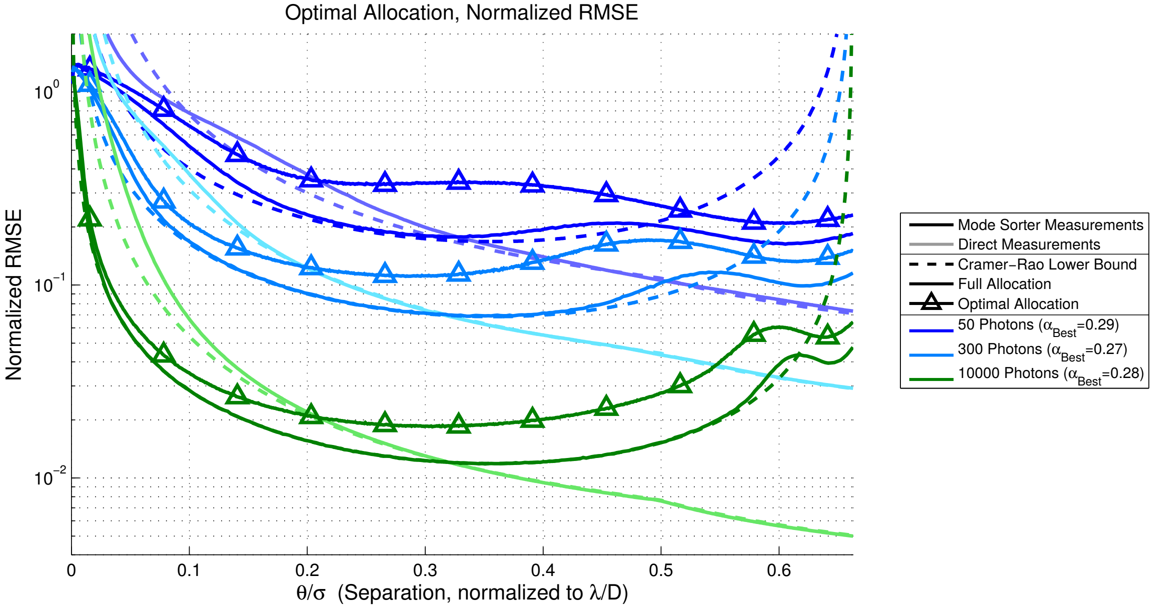


Figure 5.18: Comparison of the normalized RMSE performance with the optimum allocation strategy. The input fields are projected against the 1st mode

5.7 Conclusion and Future Work

Rayleigh's two point-sources resolution criterion describes a visual limit to the imaging resolution of a traditional imager employing a focal plane array. This visual limit is indeed not fundamental and arises in part due to the specific choice of optical measurement with a FPA in the image plane. Given the specific task of estimating the angular separation between two indistinguishable, incoherent, quasi-monochromatic point-sources, we have shown quantitatively, using Fisher's information measure that a conventional imager will perform poorly at narrow angles below Rayleigh's limit.

However as discussed earlier, Tsang *et al.* and others have firmly demonstrated that while the conventional imager is indeed subject to Rayleigh's limit for this mission, it is possible to devise an alternate linear field preprocessing (projection) passive measurement strategy which does not suffer from Rayleigh's curse. This alternate measurement performs a linear decomposition of an incident optical field in the image plane onto an orthonormal basis of spatial modes, collecting a few modal measurements with shot-noise limited detectors. Our analysis has demonstrated that the proposed architecture can outperform the traditional imager, especially in the sub-Rayleigh regime which can greatly impact many applications such as long standoff imaging, microscopy, astronomy and high-precision metrology.

Our specific contribution in this area include the extension of the original work for a Gaussian apodized pupil to a hard aperture more relevant to optical systems. We have also generalized the BinSPADE measurement design to an arbitrary aperture function, including complex-valued and asymmetrical ones. This generalization has shown that any known optical phase aberrations can be perfectly compensated by such a modal measurement approach and that greater spatial variability in the system PSF increases the amount of relevant information. Beyond the Fisher information and Cramer-Rao bound analysis, we also analyzed the RMSE performance via Monte-Carlo simulations. Finally, we discussed the performance of two candidate measurement architectures of the mode-decomposition approach for future experimental implementations.

Beyond the current work, several challenges remain in translating this alternate measurement method to real-world applications. Notably, some of the main challenges include : (a) sensitivity to centroid misalignment, (b) unequal point-source strength and (c) gaps between the ideal physical forward model and the actual ASF. Future work in this area would include adding hypothesis testing so that the device can determine if there are two sources or only one point-source, such as separately investigated by Helstrom [82] and Krovi *et al.* [86], and the ability to image coherent and incoherent extended objects.

Chapter 6

Conclusions And Future Work

In this dissertation, we have investigated the design, optimization and implementation of several computational imaging systems. In all cases, these novel imaging systems outperform conventional imaging system designs for a given specific task. We began with a joint-design approach to the extended the depth of field imaging of a camera operating at a fast f/#. Then we described a scalable compressive imaging architecture for image formation and target detection/classification tasks. Finally, we analyzed a mode-sorter imaging architecture for the task of estimating the angular separation of two incoherent quasi-monochromatic point-sources and extended the binary mode design to arbitrary aperture. We observed that these non-traditional system architectures often reveal counter-intuitive aspects of these particular imaging tasks. For instance, the EDoF architecture shows that by introducing aberration in the optical train, and thus lowering Strehl's ratio, we can render the system modulation transfer function insensitive to defocus albeit at the cost of lower modulation magnitude, which can be compensated by processing. The compressive imaging architecture demonstrates that Nyquist sampling requirement is overly restrictive when considering sparse/compressible image (such as natural scenes) and therefore it is possible to acquire high resolution images or to perform target detection/classification from a small number of measurements, *e.g.* 4x to 8x fewer than Nyquist sampling rate. Finally, the analysis of a mode-sorter imaging architecture shows that Rayleigh's limit to the resolution of an incoherent imaging system derives solely from the choice of measurement (*i.e.* focal plane array). While we observed that these computational imaging architectures have significant advantages over traditional imaging system designs, they are subject to set of unique challenges with respect to design and implementation. In this work, we have addressed several system designs relating to scalability aspects of computational imaging architectures. While scalability of traditional imagers typically refers to their capability at handling larger spatial-spectral-temporal bandwidth products, in the case of computational imaging architectures the scalability extends to their ability to manage more complex scene models, system design, optimization framework capable of exploring vast parameter spaces (optics and processing parameters) and to implement systems, including algorithms, which can perform robust image or measurements exploitation in real-time.

In chapter 2, we proposed a joint design framework for the EDoF computational imaging architecture. The system design metric provides a measure of the overall system modulation transfer function which includes both the optical and processing transfer functions. Hence, the optimization framework can simultaneously optimize the optical and processing system design parameters. We implemented this joint-design framework to optimize the parameters of an optical phase mask (Zernike polynomials coefficients) and a post-processing algorithm for a F/1 long-wave infrared imaging system operating at room temperature. We demonstrated that our optimized EDoF design achieves a larger depth of field than the traditional imaging design with minimum artifacts and has lower reconstruction noise-gain (< -1dB) than other state-of-the-art EDoF imager designs such as cubic and trefoil phase masks. To verify the pertinence of our EDoF imager design we fabricated the optimized Zernike phase mask (a free-form surface) and implemented it in an EDoF imager prototype that generated high fidelity images. In the future, we can adapt this framework and the computational imaging architecture to incorporate field aberrations and increase the constrained depth-of-focus in small form-factor camera modules (*e.g.* to compensate the field curvature).

In chapter 3, we described the implementation of a scalable and programmable compressive imager testbed which leverages the sparsity of natural images to acquire 4x to 8x fewer measurements than a traditional imager operating at Nyquist sampling. To achieve this measurement count reduction, the imaging system modulates the optical image with a passive spatial light modulator and measures the integral optical signal. Specifically, our proposed scalable architecture decomposes the field-of-view into contiguous blocks, where each is projected, measured and reconstructed in parallel so that this implementation scales to wide FOV at high spatial resolution. For the reconstruction algorithm, we have designed a piecewise-linear and non-iterative algorithm which minimizes the mean square error of the reconstructed image. This parallelizable algorithm is implemented on graphics processing units and is capable of reconstructing megapixel-scale images in real-time. Furthermore, we have implemented mitigation techniques and an automated calibration procedure for this compressive imaging testbed so as to minimize errors between the ideal forward model and the actual physical measurement implementation. Using this imaging testbed we were able to demonstrate the superior imaging performance of information-optimal projection patterns over several random projection patterns typically employed in compressive imaging systems. Specifically, we have shown that the former provides a reconstruction PSNR gain of about 6dB over the latter at 4x compression and 20dB peak measurement SNR, and up to 3dB at 8x compression. In the future, we can replace the static measurement design by an adaptive strategy which benefits greatly the measurement compression ratio. With this programmable testbed we can also imagine implementing a foveated imager with on-demand high-resolution streams from selected regions of interest.

We subsequently described in chapter 4 how this scalable programmable imaging architecture can be adapted to an automatic target detection and classification task. We showed that using appropriately designed compressive measurements we can avoid the need for intermediate image reconstruction and that we can implement the exploitation algorithm that operates directly on compressed measurements. We have used this compressive measurement strategy in conjunction to the block wise measurements design to detect target vehicles embedded in a natural cluttered scene. Similarly to the image formation task, this compressive measurement approach is naturally scalable to high space-bandwidth product (*i.e.* a wide field of view and high-resolution). Due to lower measurement count this approach requires a fraction of the readout bandwidth and power consumption budget of a conventional imaging system which employs a high-resolution FPA. To develop a statistical model of the underlying manifold describing real-world scenes, we have employed large Gaussian mixtures that incorporate the target and background variability. Using this scene model, we have developed a compressive measurement design employing Cauchy-Schwarz mutual information (CSMI). This information-theoretic metric provides a measure of the separation between the different target and background classes which ultimately determines the target detection/classification performance. Furthermore, we were able to derive an analytical upper bound to the probability of error based on CSMI. Using the compressive testbed, we verified the experimental performance of our information optimized projection patterns which achieve better target classification accuracy relative to random and PCA-secants projection designs. For example, our design achieved an accuracy of 95% correct target detection and classification at 0dB measurement SNR and a compression rate of 64x, and an accuracy greater than 97% correct target detection and classification at 10dB measurement SNR and a compression rate of 42x. In the future, we will extend our scene models to cover more scene variability, including for example occlusions and change in lighting conditions. We can also imagine adding to this compressive target recognition sensor the capability of autonomously acquiring images of the regions of interests marked as containing targets.

Finally in chapter 5, we have first quantified the performance of a conventional incoherent and diffraction-limited imager with respect to the angular separation of two incoherent and quasi-monochromatic point sources. Consistently with Rayleigh's two point-sources resolvability criterion, we have found that the conventional imaging architecture employing an ideal FPA is essentially insensitive to small angle values, *i.e.* below Rayleigh's limit. However, motivated by prior work by Tsang *et al.*, we were able to prove that the spatial mode-sorting imaging architecture, which performs a linear decomposition the optical field in the image plane, is not subject to the same Rayleigh limit by deriving the underlying Fisher Information, and the associated Cramer-Rao lower bound of the mode measurement. In particular, we were able to show that the optimal mode basis is matched to the aperture of the imaging system. For example, the Hermite-Gauss spatial mode basis for a Gaussian-apodized pupil and the Sinc-Bessel mode basis for a hard aperture. Given this observation, we were able to construct a generalized mode measurement basis for an arbitrary pupil function. We have analyzed the energy and information content of this mode-sorter architecture and showed that, for small angular separations the 0^{th} mode of the basis collects nearly all the energy and the 1^{st} mode collects nearly all information. Consequently, we devised a simple binary mode measurement architecture and proved that it significantly outperforms a traditional imaging system with FPA measurements. In a high flux regime, we found a gain of about +4dB at 10% separation and +8dB at 5% separation on the scale of the Rayleigh limit ($i.e.\ \lambda/D$). Finally, we described and analyzed the theoretical performance of two candidate implementations of the mode-sorting architecture : (a) a two-mode projections based on a Mach-Zehnder interferometer and (b) a single-mode projection based on a self-referencing volume hologram. In the future, we will improve this general mode measurement approach to test if one or two sources are visible. We will also increase the resilience of the architecture against the varying location of the point-sources pair centroid. And finally, we will adapt it so as to measure physical characteristics of small extended objects.

Appendix A

Extended Depth of Field Imaging : Supplementary Figures

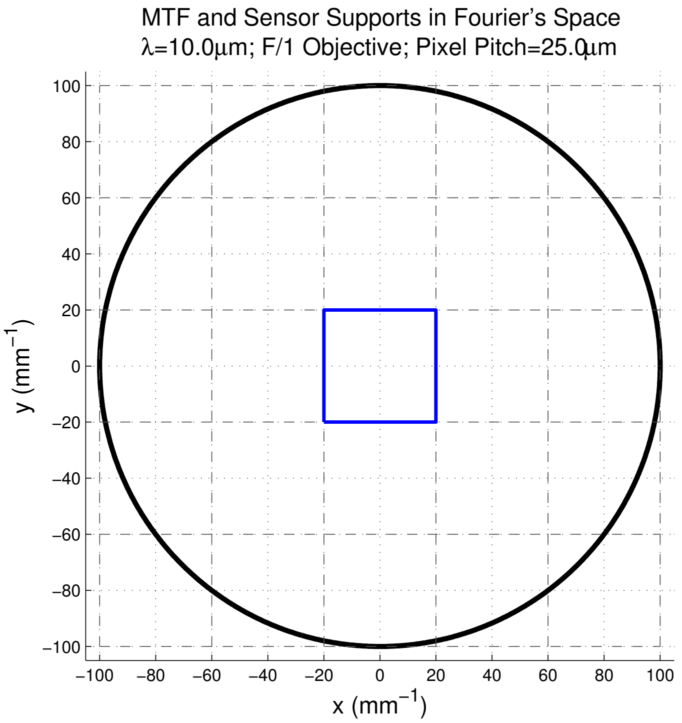


Figure A.1: Visual comparison between the optical cut-off frequency (100mm^{-1}, black circle) for the considered lens, and the sensor Nyquist frequency (20mm^{-1}, blue square).

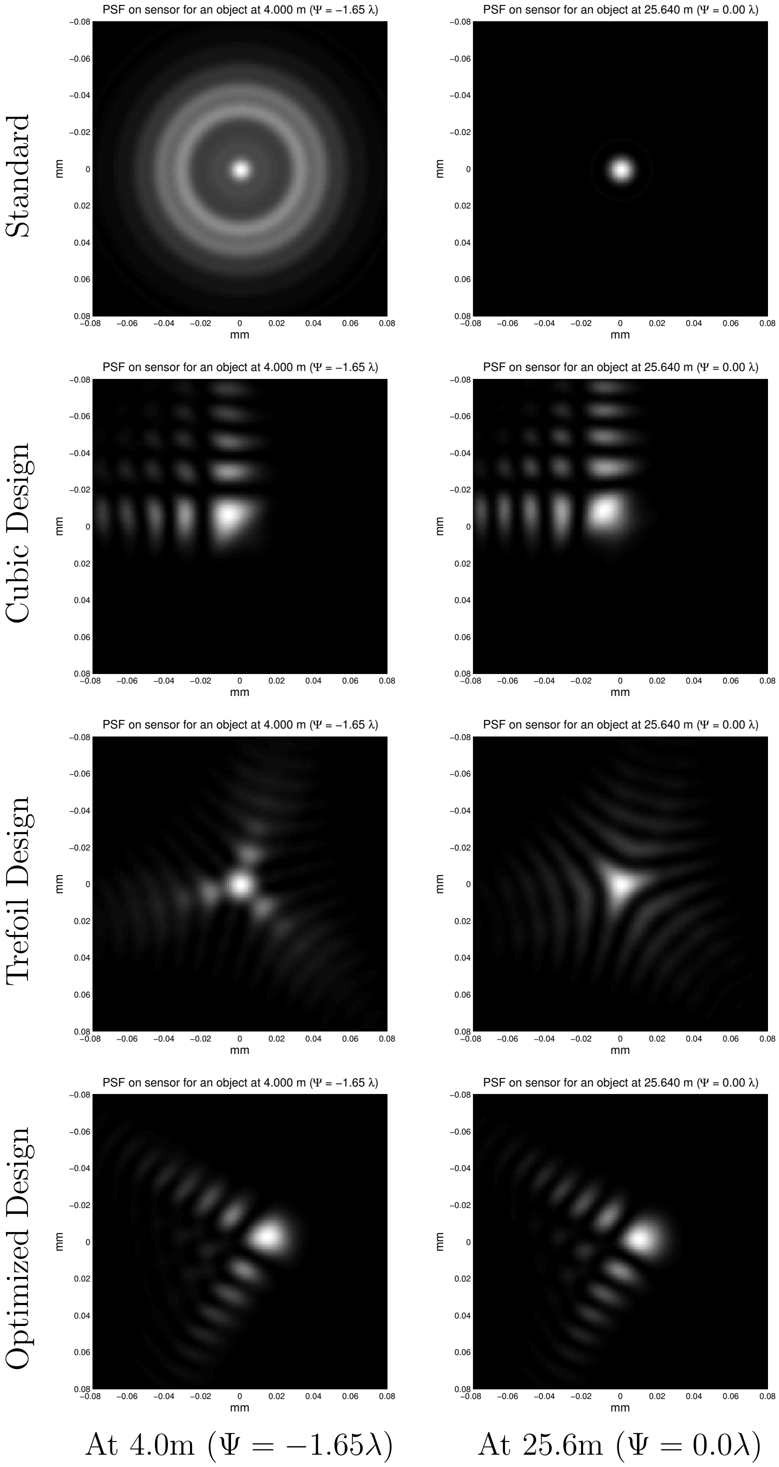


Figure A.2: Point spread functions for the standard, cubic, trefoil and optimized designs at the edge of the depth range (4.0m) and the hyperfocal distance (25.6m). We assume a perfect 25mm F/1 objective operating at a wavelength $\lambda = 10\mu$m. The figures are independently normalized to a unit peak intensity for visibility.

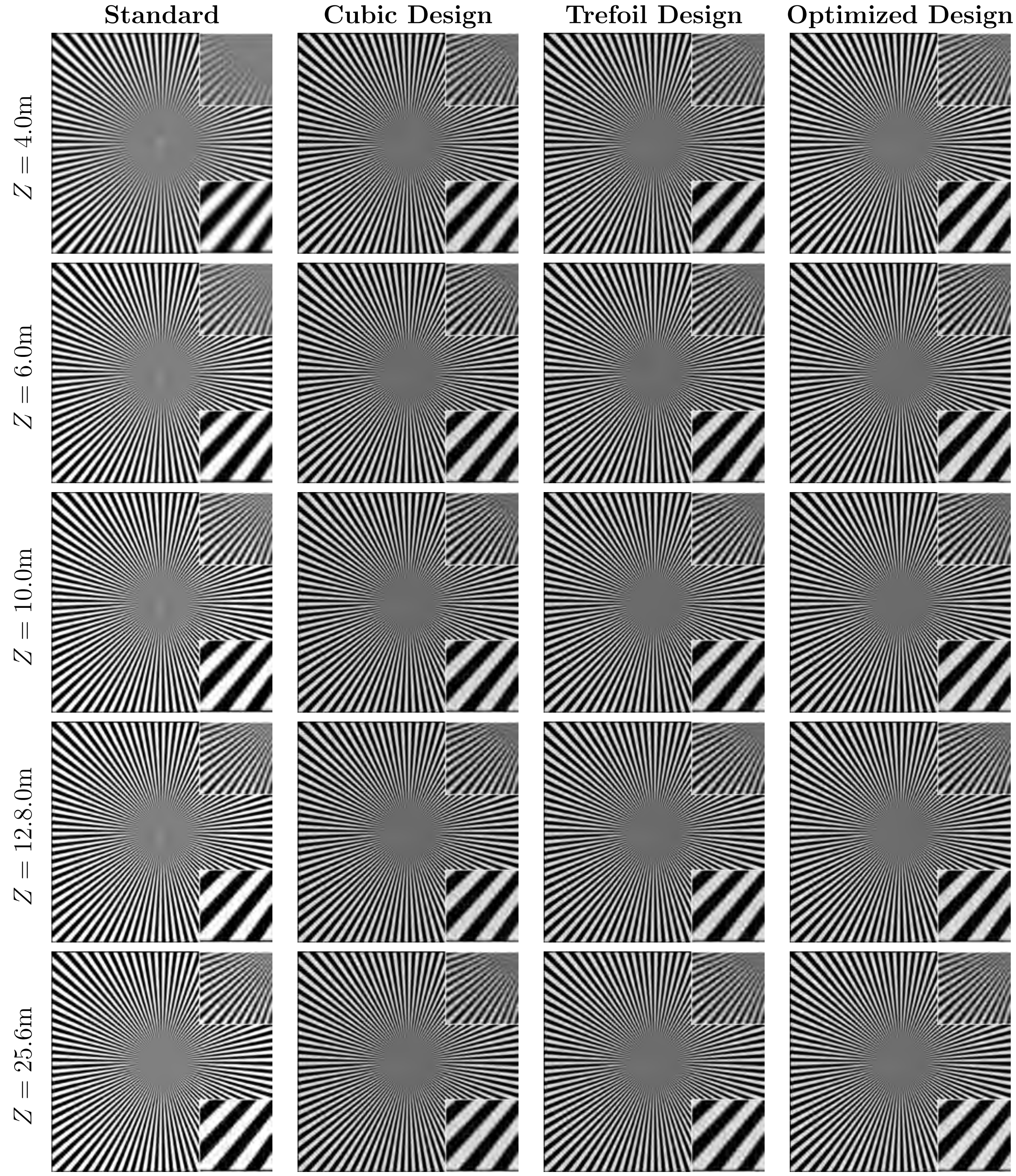


Figure A.3: Simulation of images over the Sensor OTF support and after reconstruction, for a SNR of 25dB. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

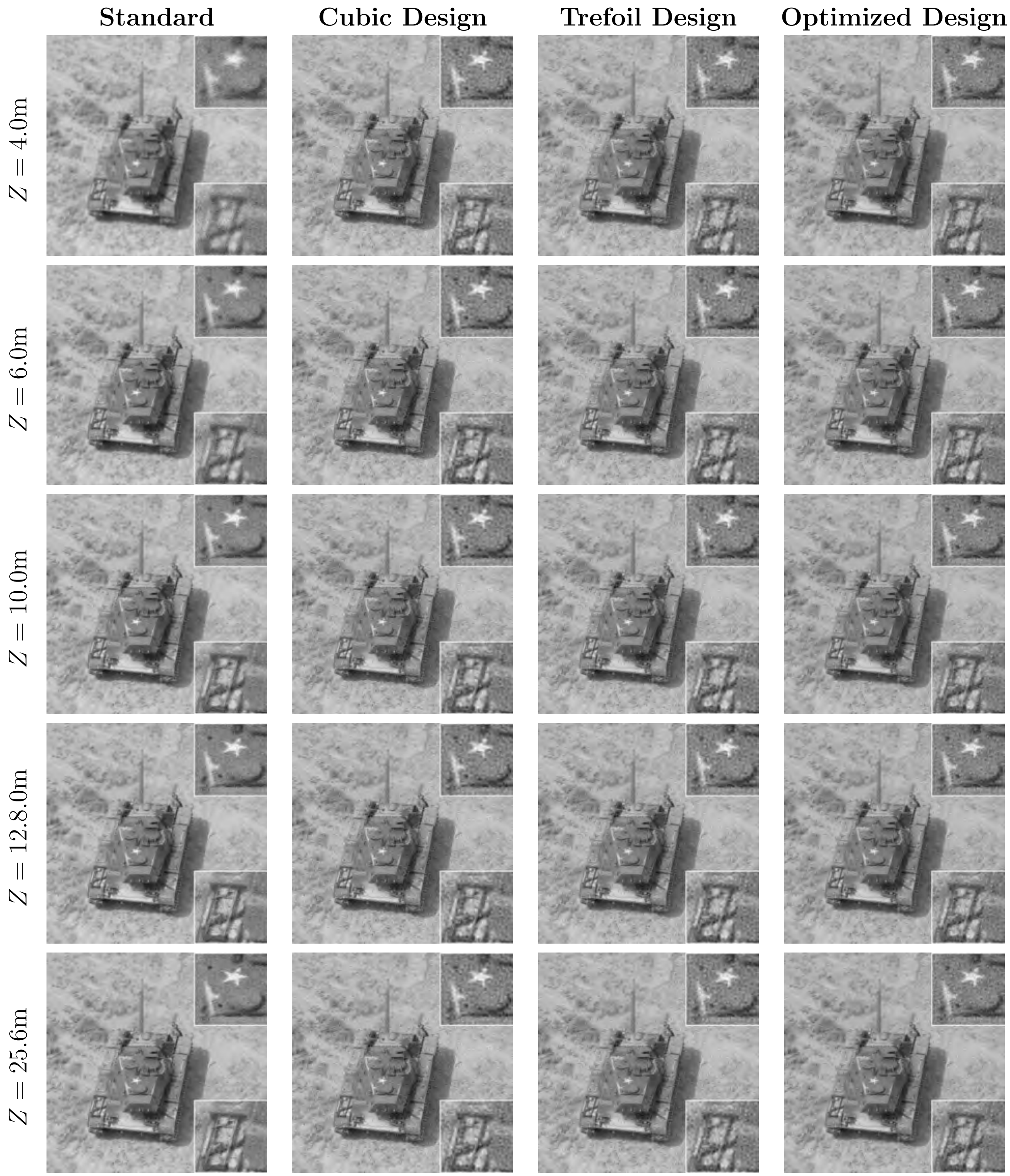


Figure A.4: Simulation of images over the Sensor OTF support and after reconstruction, for a SNR of 25dB. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

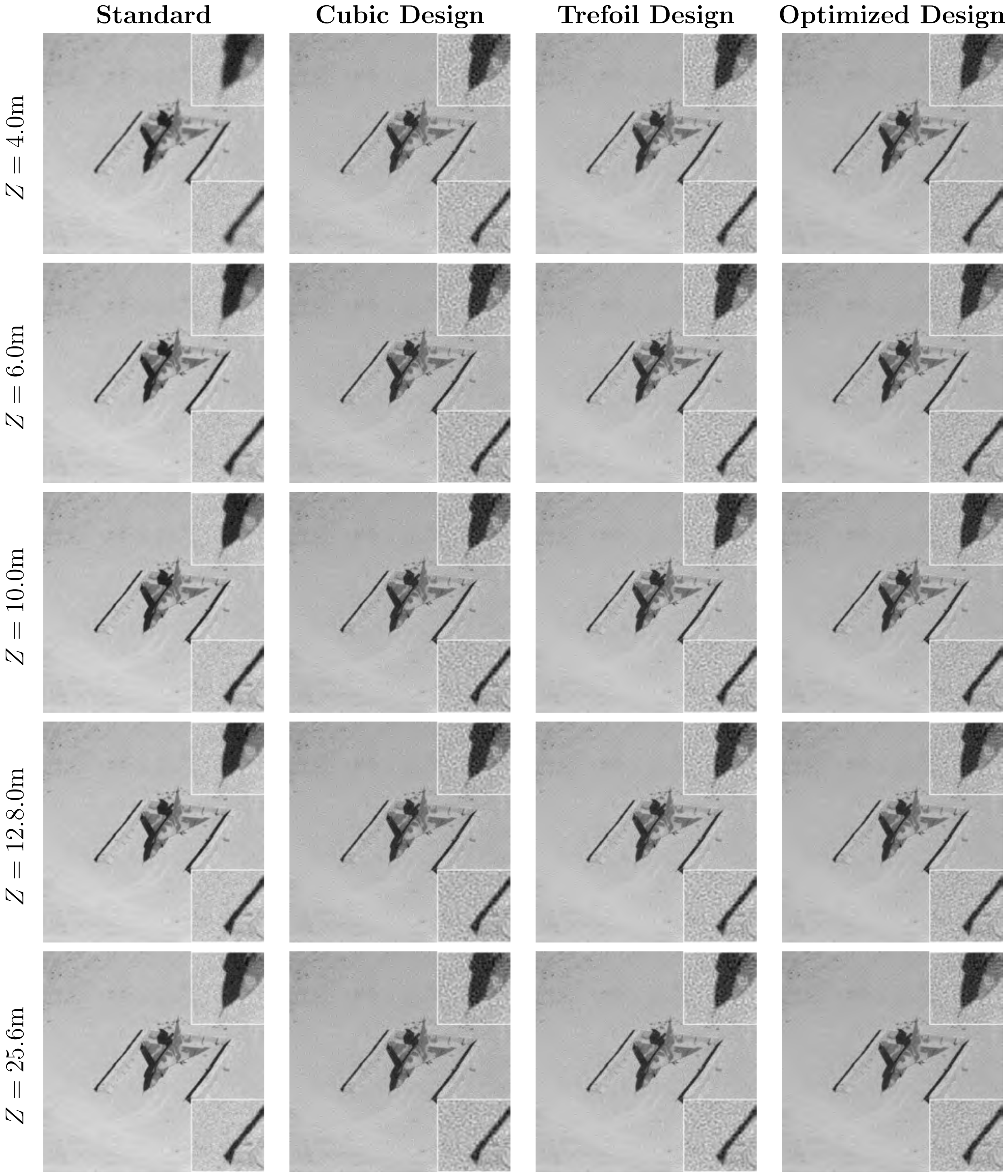


Figure A.5: Simulation of images over the Sensor OTF support and after reconstruction, for a SNR of 25dB. The distances are, from the top to bottom row : $4.0m$ ($\Psi = -1.65\lambda$), $6.0m$ ($\Psi = -1.00\lambda$), $10.0m$ ($\Psi = -0.67\lambda$), $12.8m$ ($\Psi = -0.31\lambda$) and $25.6m$ ($\Psi = 0$, the hyperfocal distance).

Appendix B

Scalable Compressive Target Detection and Classification Sensor : Additional Derivations

B.1 Derivation of the CSMI Expression

First, the definition of Cauchy-Schwarz mutual information is recalled in the case of two random variables associated by the measurement process : C denotes the N_c class label ($C \in \mathcal{C} = [\![1; N_c]\!]$) and $\boldsymbol{G}$ a real valued vector in $\mathbb{R}^{EM}$ representing the measurements made by the system. As one of the variable is discrete, we write with $p_{C,\boldsymbol{G}}$ their joint distribution and $p_C p_{\boldsymbol{G}}$ the product of their marginals, and using Bayes rule :

$$\mathcal{I}_{\mathrm{CS}}(C,\boldsymbol{G}) = -\ln\left[\frac{\sum_{c\in\mathcal{C}}\int_{\mathbb{R}^{EM}} p_{C,\boldsymbol{G}}(c,\boldsymbol{g})p_C(c)p_{\boldsymbol{G}}(\boldsymbol{g})d^{EM}\boldsymbol{g}}{\sqrt{\left(\sum_{c\in\mathcal{C}}\int_{\mathbb{R}^{EM}} p_{C,\boldsymbol{G}}(c,\boldsymbol{g})^2 d^{EM}\boldsymbol{g}\right)\left(\sum_{c\in\mathcal{C}}\int_{\mathbb{R}^{EM}} p_C(c)^2 p_{\boldsymbol{G}}(\boldsymbol{g})^2 d^{EM}\boldsymbol{g}\right)}}\right] \tag{B.1}$$

$$= -\ln\left[\frac{\sum\sum_{i,j\in\mathcal{C}} p_C(i)^2 p_C(j) V_{i,j}}{\sqrt{\left(\sum_{i\in\mathcal{C}} p_C(i)^2\right)\left(\sum_{i\in\mathcal{C}} p_C(i)^2 V_{i,i}\right)\left(\sum\sum_{i,j\in\mathcal{C}} p_C(c_i)p_C(c_j)V_{i,j}\right)}}\right]. \tag{B.2}$$

Where the symmetric matrix $V \in \mathbb{R}^{+N_c\times N_c}$ is directly indexed by the class labels and results from the likelihoods overlap integrals, *i.e.* : the overlap coefficients are $V_{i,j} = \int_{\mathbb{R}^{EM}} p(\boldsymbol{g}|C=i)p(\boldsymbol{g}|C=j)d^{EM}\boldsymbol{g}$. One can also remark that these coefficients are obtained as the results of dot-products between the likelihoods and, as such, obey both the triangular and Cauchy-Schwarz inequalities.

In this work, all the likelihoods $p(\boldsymbol{g}|C)$ are expressed as mixtures of Gaussians : for instance, the mixture for the class label c_i contains O_i components, each with a weight $\omega_{i,k}$ (with the normalization : $\sum_{k=1}^{O_i}\omega_{i,k} = 1$), a mean $\boldsymbol{s}_{i,k}$ (in $\mathbb{R}^{EM}$) and a covariance $S_{i,k}$ (in $\mathbb{R}^{EM\times EM}$) :

$$\begin{aligned} p(\boldsymbol{g}|C=c_i) &= \sum_{k=1}^{O_i}\omega_{i,k}\mathcal{N}(\boldsymbol{g}|\boldsymbol{s}_{i,k},S_{i,k}) \\ &= \sum_{k=1}^{O_i}\omega_{i,k}\frac{|S_{i,k}|^{-\frac{1}{2}}}{\sqrt{2\pi}^{EM}}\exp\left(-\frac{1}{2}(\boldsymbol{g}-\boldsymbol{s}_{i,k})^{\intercal}S_{i,k}^{-1}(\boldsymbol{g}-\boldsymbol{s}_{i,k})\right). \end{aligned} \tag{B.3}$$

With those likelihoods, the overlap coefficients can be written in closed form using a simple relation on the integral of a product of multivariate Gaussians : given two multivariate Gaussians of respective means $\boldsymbol{a}$, $\boldsymbol{b}$, and covariances A, B, one can verify the following : $\int_{\mathbb{R}^{EM}}\mathcal{N}(\boldsymbol{x}|\boldsymbol{a},A)\mathcal{N}(\boldsymbol{x}|\boldsymbol{b},B)\ d^{EM}\boldsymbol{x} = \mathcal{N}(\boldsymbol{a}-\boldsymbol{b}|\boldsymbol{0},A+B)$. Using this property, the CSMI expression can be expanded analytically into three main terms, detailing the components interactions within and between the classes, by developing the overlap coefficients :

$$V_{i,j} = \sum_{k=1}^{O_i} \sum_{l=1}^{O_j} \omega_{i,k}\omega_{j,l}\mathcal{N}\left(\boldsymbol{s}_{i,k} - \boldsymbol{s}_{j,l}|\boldsymbol{0}, S_{i,k} + S_{j,l}\right). \tag{B.4}$$

Thus, the value of CSMI is directly related to the sum of the overlap matrix V.

Until now, the expression given in Equation B.2 is purely generic and we can continue by injecting our specific statistical model and our linear measurement models. The first is developed from data acquired in the direct domain (from conventional cameras, noted $\boldsymbol{f} \in \mathbb{R}^{EM}$) and must be transformed by the second, consisting of a linear projection operator P and additive white Gaussian noise $\boldsymbol{z}$ of variance σ_n^2. The compression of the likelihoods from the direct space is expressed as follow :

$$\begin{aligned} p(\boldsymbol{f}|C=i) &= \sum_{k=1}^{O_i} \omega_{i,k}\mathcal{N}(\boldsymbol{g}|\boldsymbol{t}_{i,k}, T_{i,k}) \\ \underset{\boldsymbol{g}=(I_E\otimes P)\boldsymbol{f}+\boldsymbol{z}}{\Rightarrow} p(\boldsymbol{g}|C=i) &= \sum_{k=1}^{O_i} \omega_{i,k}\mathcal{N}\left(\boldsymbol{g}|\boldsymbol{s}_{i,k} I_E \otimes P)\boldsymbol{t}_{i,k},\ S_{i,k} = \right. \\ &\qquad \left.((I_E \otimes P)T_{i,k}(I_E \otimes P)^\intercal + \sigma_n^2 I_{EM}\right). \end{aligned} \tag{B.5}$$

In order to both limit the computational complexity and produce and upper-bound on the probability of error, all the direct-space covariances $T_{i,j}$ are set to be spherical, *i.e.* $\forall i,k,\ T_{i,k} = \sigma^2 I_{EN}$ where the scalar variance σ^2 is obtained after training the statistical model. Using the properties of the Kronecker product for matrices of matching size, the generic overlap coefficients given in Equation B.4 become :

$$V_{i,j} = \sum_{k=1}^{O_i} \sum_{l=1}^{O_j} \omega_{i,k}\omega_{j,l}\mathcal{N}\left(\boldsymbol{s}_{i,k} - \boldsymbol{s}_{j,l}|\boldsymbol{0}, \left(I_E \otimes 2(\sigma^2 PP^\intercal + \sigma_n^2 I_M)\right)\right), \tag{B.6}$$

$$V^*_{i,j} = \sum_{k=1}^{O_i} \sum_{l=1}^{O_j} \omega_{i,k}\omega_{j,l} \exp\left(-\frac{1}{4}\operatorname{Tr}\left[(U_{i,k} - U_{j,l})^\intercal P^\intercal (\sigma^2 PP^\intercal + \sigma_n^2 I_M)^{-1} P (U_{i,k} - U_{j,l})\right]\right). \tag{B.7}$$

Where $V^*_{i,j}$ is the denormalized overlap coefficient obtained after removing the superfluous distribution normalization (for the computation of CSMI) and replacing the direct-space samples $t_{i,k}$ by their matrix folded (or re-arranged) counterparts : $U_{i,k} \in \mathbb{R}^{N\times E}$, consisting of E blocks of dimensions N.

B.2 Derivation of the Gradient Expression

It is now simple, from this expression, to device the gradient of the CSMI metric with respect to the projection matrix P. For this, we will rely extensively on the chain rule as well as matrix and Jacobian identities. One can write for the denormalized overlap coefficients, with the definition of the compressed covariance $\Sigma = \sigma^2 P P^\intercal + \sigma_n^2 I_M$:

$$\begin{aligned}\boldsymbol{\nabla}\left(V^*_{i,j}(P)\right) = -\frac{1}{2}\sum_{k=1}^{O_i}\sum_{l=1}^{O_j}\omega_{i,k}\omega_{j,l}\exp\left(-\frac{1}{4}\operatorname{Tr}\left[(U_{i,k}-U_{j,l})^\intercal P^\intercal \Sigma^{-1} P (U_{i,k}-U_{j,l})\right]\right)\\ \times\, \Sigma^{-1}P(U_{i,k}-U_{j,l})\left[(U_{i,k}-U_{j,l}) - \sigma^2 P^\intercal \Sigma^{-1} P(U_{i,k}-U_{j,l})\right]^\intercal . \qquad \text{(B.8)}\end{aligned}$$

Finally, the gradient of the overlap coefficients is used to develop the expression of interest :

$$\begin{aligned}\boldsymbol{\nabla}\left(\mathcal{I}_{\mathrm{CS}}(P)\right) = \frac{1}{2}\frac{\sum_{c\in\mathcal{C}} p_C(c)^2 \boldsymbol{\nabla}\left(V^*_{c,c}(P)\right)}{\sum_{c\in\mathcal{C}} p_C(c)^2 V^*_{c,c}} + \frac{1}{2}\frac{\sum\sum_{i,j\in\mathcal{C}} p_C(i)p_C(j)\boldsymbol{\nabla}\left(V^*_{i,j}(P)\right)}{\sum\sum_{i,j\in\mathcal{C}} p_C(i)p_C(j)V^*_{i,j}}\\ - \frac{\sum\sum_{i,j\in\mathcal{C}} p_C(i)^2 p_C(j)\boldsymbol{\nabla}\left(V^*_{i,j}(P)\right)}{\sum\sum_{i,j\in\mathcal{C}} p_C(i)^2 p_C(j) V^*_{i,j}}. \qquad \text{(B.9)}\end{aligned}$$

B.3 Derivation of the Upper-Bound on the Probability of Error

First, the probability of error is defined as the integral of the likelihoods over the regions where the ideal (MAP) classifier would not output their label and is written :

$$\mathrm{P_e} = 1 - \int_{\mathbb{R}^{EM}} \max_{i\in\mathcal{C}} \{p(\boldsymbol{g}|C=i)\} \;\, d^{EM}\boldsymbol{g}. \tag{B.10}$$

Here, it is assumed that the N_c classes are equiprobable ($p_C(c) = 1/N_c$) and also that all the likelihoods are mixtures of Gaussians, as defined in Equation B.3. This expression is then developed by using the positivity of both weights and statistical distributions, in order to access simple pairwise interactions at the cost of producing an upper-bound. We have :

$$\begin{aligned}\mathrm{P_e} &\leq \frac{1}{N_c}\sum_{i>j\in\mathcal{C}}\sum\sum_{k=1}^{O_i}\sum_{l=1}^{O_j}\max\{\omega_{i,k},\omega_{j,l}\}\underbrace{\int_{\mathbb{R}^M}\min\{\mathcal{N}(\boldsymbol{g}|\boldsymbol{s}_{i,k},S_{i,k}),\mathcal{N}(\boldsymbol{g}|\boldsymbol{s}_{j,l},S_{j,l})\}\;d^M\boldsymbol{g}}_{=2P_{e:i,k,j,l}}\\ &\leq \frac{2}{N_c\min\{\omega_{i,k}\}}\sum_{i>j\in\mathcal{C}}\sum\sum_{k=1}^{O_i}\sum_{l=1}^{O_j}\omega_{i,k}\omega_{j,l}P_{e:i,k;j,l}.\end{aligned} \tag{B.11}$$

From Equation B.11, a generalized Bhattacharyya bound $\mathrm{U_{Pe,Bc}}$ can be expressed as :

$$\mathrm{P_e} \leq \frac{1}{N_c\min\{\omega_{i,k}\}}\sum_{i,j\in\mathcal{C},j>i}\sum\sum_{k=1}^{O_i}\sum_{l=1}^{O_j}\omega_{i,k}\omega_{j,l}B_{c:i,k;j,l} = \mathrm{U_{Pe,Bc}} \tag{B.12}$$

$$\begin{aligned}B_{c:i,k;j,l} &= \int_{\mathbb{R}^{EM}}\sqrt{\mathcal{N}(\boldsymbol{g}|\boldsymbol{s}_{i,k},S_{i,k})\mathcal{N}(\boldsymbol{g}|\boldsymbol{s}_{j,l},S_{j,l})}\;d^{EM}\boldsymbol{g}\\ &= (8\pi)^{\frac{M}{2}}|S_{i,k}S_{j,l}|^{\frac{1}{4}}\mathcal{N}(\boldsymbol{s}_{i,k}-\boldsymbol{s}_{j,l}|\boldsymbol{0},2(S_{i,k}+S_{j,l})).\end{aligned} \tag{B.13}$$

This inequality can be quadratically curved to relate the Bhattacharyya pairwise coefficients to the overlap coefficients used in CSMI :

$$B^2_{c:i,k;j,l} = (8\pi)^{\frac{M}{2}} \sqrt{\frac{|S_{i,k} S_{j,l}|}{|S_{i,k} + S_{j,l}|}} \mathcal{N}(\boldsymbol{s}_{i,k} - \boldsymbol{s}_{j,l} | \mathbf{0}, S_{i,k} + S_{j,l}), \tag{B.14}$$

$$\frac{4 \min\{\omega^2_{i,k}\}}{(N_c - 1)^2} \, \mathrm{P_e}^2 \leq \sum\sum_{i>j \in \mathcal{C}} \sum_{k=1}^{O_i} \sum_{l=1}^{O_j} \frac{2 \omega_{i,k} \omega_{j,l}}{N_c (N_c - 1)} B^2_{c:i,k;j,l}. \tag{B.15}$$

For the expression of CSMI, Equation B.2 can be adapted to the case of equiprobable classes as :

$$\mathcal{I}_{\mathrm{CS}}(C, \boldsymbol{G}) = \frac{1}{2} \ln(N_c) - \frac{1}{2} \ln \left[1 + 2 \frac{\sum\sum_{i>j \in \mathcal{C}} V_{i,j}}{\sum_{i \in \mathcal{C}} V_{i,i}} \right]. \tag{B.16}$$

It immediately appears that the following upper and lower bounds apply : $0 \leq \mathcal{I}_{\mathrm{CS}} \leq \ln(N_c)/2$. And, from the definition of the overlap coefficients :

$$V_{i,i} \leq (2\pi)^{-\frac{M}{2}} \max_{k,l \leq O_i} \left\{ |S_{i,k} + S_{i,l}|^{-\frac{1}{2}} \right\}, \tag{B.17}$$

$$V_{i,j} = \sum_{k=1}^{O_i} \sum_{l=1}^{O_j} \omega_{i,k} \omega_{j,l} (8\pi)^{-\frac{M}{2}} \sqrt{\frac{|S_{i,k} + S_{j,l}|}{|S_{i,k} S_{j,l}|}} B^2_{c:i,k;j,l}. \tag{B.18}$$

Thus, the probability of error is related to the overlap coefficients expression via the inequality :

$$\frac{\sum\sum_{i>j \in \mathcal{C}} V_{i,j}}{\sum_{i \in \mathcal{C}} V_{i,i}} \geq \frac{2 \min\{\omega^2\}}{(N_c - 1)} \frac{\min\{|S_{i,k} + S_{j,l}|\}}{2^N \sqrt{\max\{|S_{i,k} S_{j,l}|\}}} \, \mathrm{P_e}^2. \tag{B.19}$$

Finally, we conclude :

$$\mathrm{P_e} \leq \mathrm{U_{Pe}} = \frac{1}{2 \min\{\omega_{i,k}\}} \sqrt{(N_c - 1) K_{\mathrm{Sat}} \left[\exp(\ln(N_c) - 2\, \mathcal{I}_{\mathrm{CS}}) - 1 \right]}, \tag{B.20}$$

$$\text{with : } K_{\mathrm{Sat}} = \frac{\sqrt{\max\{|S_{i,k} S_{j,l}|\}}}{\min\{|(S_{i,k} + S_{j,l})/2|\}}.$$

One can remark immediately that in the particular case where all the covariances $S_{i,k}$ are identical, the saturation factor is then $K_{\text{Sat}} = 1$. Consequently, CSMI only depends on the projector P in the upper-bound expression. Its gradient can simply be written as :

$$\boldsymbol{\nabla}\left(\mathrm{U}_{\mathrm{Pe}}(P)\right) = \frac{\sqrt{N_c - 1}}{2\min\{\omega_{i,k}\}} f\left[\exp(\ln(N_c) - 2\,\mathcal{I}_{\mathrm{CS}}(P))\right] \boldsymbol{\nabla}\left(\mathcal{I}_{\mathrm{CS}}(P)\right), \qquad \text{(B.21)}$$
$$\text{With : } f : x \mapsto \frac{-x}{\sqrt{x-1}}.$$

As a final note, we remark that Equation B.12 on the Bhattacharyya information (and such as presented in [56, 70]) provides a tighter upper bound on the probability of error than CSMI : *i.e.* $\mathrm{P_e} \leq \mathrm{U_{Pe,Bc}} \leq \mathrm{U_{Pe}}$. However, this metric only consider the distance separating components of different classes and not the distance between components of the same class, as previously illustrated for the CSMI metric in Figure 4.8. Hence, optimizing the Bhattacharyya information is similar to *pushing away* components belonging to different class likelihoods but not *pulling together* components belonging to the same class such as it is the case for CSMI.

Appendix C

Two Point-Sources Resolution Measurement Design : Additional Derivations

Proof 1 : Fisher Information for a Poisson corrupted process. Let f be a function $\mathbb{R} \rightarrow \mathbb{R}^+$ modeling the output of a process on a variable θ that is corrupted by Poisson noise :

$$p_f(y|\theta) = \frac{f(\theta)^y \exp(-f(\theta))}{y!}, \tag{C.1}$$

such that $\ln(p_f(y|\theta))$ is twice differentiable with respect to θ. We note f' and f'', respectively its first and second derivatives. Then we can write the corresponding Fisher Information as

:

$$\begin{aligned}
\mathcal{I}(\theta) &= -\,\mathbf{E}_y\left[\frac{\partial^2}{\partial\theta^2}\ln(p_f(y|\theta))\middle|\,\theta\right] && \text{(C.2)}\\
&= -\frac{\mathbf{E}_y[y|\theta]}{f(\theta)}\left(f''(\theta)-\frac{f'(\theta)^2}{f(\theta)}\right)+f''(\theta) && \text{(C.3)}\\
&= \frac{f'(\theta)^2}{f(\theta)} && \text{(C.4)}
\end{aligned}$$

Proof 2 : Fisher Information For Multiple Independent Measurements. If we consider the output of multiple independent measurement functions $(f_0(\theta),\dots,f_Q(\theta))$, all twice differentiable with respect to θ, and their respective outputs $y_0,\dots,y_Q$, we have for the Fisher Information :

$$\begin{aligned}
\mathcal{I}(\theta) &= -\,\mathbf{E}_{y_0,\dots,y_Q}\left[\frac{\partial^2}{\partial\theta^2}\ln\left(\prod_{q=0}^{Q}p_q(y_q|\theta)\right)\middle|\,\theta\right] && \text{(C.5)}\\
&= \sum_{q=0}^{Q}\frac{f_q'(\theta)^2}{f_q(\theta)} = \sum_{q=0}^{Q}\mathcal{I}_q(\theta) && \text{(C.6)}
\end{aligned}$$

Lemma 1 : Uniform Convergence Of Series Based On Spherical Bessel Functions Of The First Kind. We consider the two following series of functions over a finite interval I of $\mathbb{R}$ containing zero, for some fixed positive integer b :

$$\begin{aligned}
\forall Q\in\mathbb{N},\ x\in I,\ A_{b,Q}(x) &= \sum_{q=0}^{Q}q^b\frac{\partial}{\partial x}j_q(x)^2 && \text{(C.7)}\\
B_{b,Q}(x) &= \sum_{q=0}^{Q}q^b\frac{\partial}{\partial x}j_q(x)j_{q+1}(x) && \text{(C.8)}
\end{aligned}$$

We have the following loose upper bound on the Spherical Bessel Functions, from their series definition :

$$\begin{aligned}
\forall q\in\mathbb{N},\ x\in\mathbb{R},\ j_q(x) &= \frac{\sqrt{\pi}}{2^{q+1}}x^q\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\Gamma\left(k+q+\frac{3}{2}\right)}\left(\frac{x}{2}\right)^{2k} && \text{(C.9)}\\
|j_q(x)| &\leq \frac{\sqrt{\pi}}{2q!}\left(\frac{|x|}{2}\right)^q\exp\left(\frac{x^2}{4}\right) && \text{(C.10)}
\end{aligned}$$

We can pick X such that, $\forall x \in I,\ |x| \leq X$ and we can test that, for the first series in Equation C.7 we have the convergence bound :

$$\begin{aligned}
&\forall P, Q \in \mathbb{N}, Q > P, \\
&|A_{b,Q}(x) - A_{b,P}(x)| = \left| \sum_{q=P+1}^{Q} 2\frac{q^{b+1}}{x} j_q(x)^2 - 2q^b j_q(x) j_{q+1}(x) \right| \qquad \text{(C.11)} \\
&\qquad \leq \pi(1 + X^2) \exp\left(\frac{X^2}{2}\right) \sum_{q=P+1}^{\infty} \frac{q^{b+1}}{2^q q!} X^{2q-1}. \qquad \text{(C.12)}
\end{aligned}$$

As the underlying series is positive, monotonic and converges (by the ratio test), we can choose a P large enough to reduce the remainder, and subsequently the bound, to any small ϵ we wish. In other words :

$$\forall \epsilon > 0, \exists P \in \mathbb{N}, \forall p, q \geq P, x \in I, |A_{b,p}(x) - A_{b,q}(x)| < \epsilon. \qquad \text{(C.13)}$$

We can thus conclude that the function series is uniformly Cauchy and converges uniformly over I.

For the second series in Equation C.8 we have:

$$\begin{aligned}
&\forall P, Q \in \mathbb{N}, Q > P, \\
&|B_{b,Q}(x) - B_{b,P}(x)| = \left| \sum_{q=P+1}^{Q} q^b j_q(x)^2 - \frac{2}{x} q^b j_q(x) j_{q+1}(x) - q^b j_{q+1}(x)^2 \right| \qquad \text{(C.14)} \\
&\qquad \leq \pi\left(1 + \frac{X^2}{2}\right) \exp\left(\frac{X^2}{2}\right) \sum_{q=P+1}^{\infty} \frac{q^b}{2^q q!} X^{2q}, \qquad \text{(C.15)}
\end{aligned}$$

where the same conclusion applies.

Finally, we consider the new set of series of functions, related respectively to $A_{b,Q}$ and

$B_{b,Q}$:

$$\forall Q, b \in \mathbb{N}, \forall x \in I,\ C_{b,Q}(x) = \sum_{q=0}^{Q} q^b j_q(x)^2 \tag{C.16}$$

$$D_{b,Q}(x) = \sum_{q=0}^{Q} q^b j_q(x) j_{q+1}(x). \tag{C.17}$$

We note that their pointwise convergence can easily be established at $x = 0$, as $\forall q > 0,\ j_q(0) = 0,\ j_0(x) = 1$ and, considering the previous results, we can apply the Differentiation Theorem to obtain their respective uniform convergence as well as the relations :

$$C_{b,Q}(x) = \frac{\partial}{\partial x} A_{b,Q}(x), \text{ and} \tag{C.18}$$

$$D_{b,Q}(x) = \frac{\partial}{\partial x} B_{b,Q}(x). \tag{C.19}$$

Proof 3 : Series of Spherical Bessel Function Of The First Kind. We will use the following two results from [78] (Equations 1.10.50 and 1.10.52), where $\mathrm{Si}(x)$ denotes the Sine Integral.

$$\sum_{q=0}^{\infty} j_q(x)^2 = \frac{\mathrm{Si}(2x)}{2x}, \text{ and} \tag{C.20}$$

$$\sum_{q=0}^{\infty} (1 + 2q) j_q(x)^2 = 1 \tag{C.21}$$

By combining them, we obtain:

$$\sum_{q=0}^{\infty} q j_q(x)^2 = \frac{1}{2}\left(1 - \frac{\mathrm{Si}(2x)}{2x}\right) \tag{C.22}$$

$$\sum_{q=0}^{\infty} j_{q+1}(x)^2 = \frac{\mathrm{Si}(2x)}{2x} - j_0(x)^2 \tag{C.23}$$

$$\sum_{q=0}^{\infty} q j_{q+1}(x)^2 = \frac{1}{2} - \frac{3\,\mathrm{Si}(2x)}{4x} + j_0(x)^2 \tag{C.24}$$

$$\sum_{q=0}^{\infty} (1+2q) j_{q+1}(x)^2 = 1 - \frac{\mathrm{Si}(2x)}{x} + j_0(x)^2 \tag{C.25}$$

Thanks to the Differentiation results of the previous lemma (C.18), we can write from the derivative of (C.20):

$$\sum_{q=0}^{\infty} j_q(x) j_{q+1}(x) = \frac{1}{2x}\left(1 - j_0(2x)\right). \tag{C.26}$$

We then find the following as the solution of a first order linear differential equation involving Equation C.22 and Equation C.24 as well as the property (C.19) and the constraint that the series is equal to 0 at $x = 0$:

$$\sum_{q=0}^{\infty} q j_q(x) j_{q+1}(x) = \frac{\mathrm{Si}(2x)}{4} + \frac{3\sin(2x)}{16x^2} - \frac{1}{2x} + \frac{\cos(2x)}{8x} \tag{C.27}$$

With this result, we can compute the derivative of Equation C.22 to express the following series and its respective shifted version :

$$\sum_{q=0}^{\infty} q^2 j_q(x)^2 = \frac{1}{8}\,\mathrm{Si}(2x)\left(\frac{1}{x} + 2x\right) + \frac{\cos(2x)}{8} + \frac{j_0(2x)}{8} - \frac{1}{2} \tag{C.28}$$

$$\sum_{q=0}^{\infty} q^2 j_{q+1}(x)^2 = \frac{1}{8}\,\mathrm{Si}(2x)\left(\frac{9}{x} + 2x\right) + \frac{\cos(2x)}{8} + \frac{j_0(2x)}{8} - j_0(x)^2 - \frac{3}{2} \tag{C.29}$$

The next series is also found as the solution of another differential equation involving

Equation C.28 and Equation C.29, with the same value constraint at $x = 0$ than previously:

$$\sum_{q=0}^{\infty} q^2 j_q(x) j_{q+1}(x) = -\frac{\mathrm{Si}(2x)}{2} - \frac{\sin(2x)}{8x^2} + \frac{x}{3} + \frac{1}{2x} - \frac{\cos(2x)}{4x} \tag{C.30}$$

And after deriving Equation C.28 we obtain:

$$\sum_{q=0}^{\infty} q^3 j_q(x)^2 = -\frac{6x^2+1}{16x}\,\mathrm{Si}(2x) - \frac{3\sin(2x) + 6x\cos(2x)}{32x} + \frac{x^2}{3} + \frac{1}{2} \tag{C.31}$$

Finally, we can combine the results of Equations C.24 to C.31 to get:

$$\sum_{q=0}^{\infty} (1+2q)\frac{q^2}{x^2} j_q(x)^2 - 2(1+2q)\frac{q}{x} j_q(x) j_{q+1}(x) + (1+2q) j_{q+1}(x)^2 = \frac{1}{3}. \tag{C.32}$$

Proof 4 : ASF Insensitivity Property. Considering any continuous and differentiable ASF $A(x)$ as well as mode $g(x)$, the measurement function m_g for two point sources separated by 2θ can be written as:

$$m_g(\theta) = \frac{1}{2\sigma^2}\left|\int_{-\infty}^{+\infty} \overline{g\left(\frac{x}{\sigma}\right)} A\left(\frac{x+\theta}{\sigma}\right)\ dx\right|^2 + \frac{1}{2\sigma^2}\left|\int_{-\infty}^{+\infty} \overline{g\left(\frac{x}{\sigma}\right)} A\left(\frac{x-\theta}{\sigma}\right)\ dx\right|^2. \tag{C.33}$$

We have for the limit of its first derivative:

$$\begin{aligned}\lim_{\theta\to 0} \frac{\partial}{\partial\theta} m_g(\theta) = \lim_{\theta\to 0}\quad & \frac{1}{\sigma^3}\mathrm{Re}\left[\int_{-\infty}^{+\infty} \overline{g\left(\frac{x}{\sigma}\right)} A'\left(\frac{x}{\sigma}\right)\ dx \overline{\int_{-\infty}^{+\infty} \overline{g\left(\frac{x}{\sigma}\right)} A\left(\frac{x}{\sigma}\right)\ dx}\right] \\ & - \frac{1}{\sigma^3}\mathrm{Re}\left[\int_{-\infty}^{+\infty} \overline{g\left(\frac{x}{\sigma}\right)} A'\left(\frac{x}{\sigma}\right)\ dx \overline{\int_{-\infty}^{+\infty} \overline{g\left(\frac{x}{\sigma}\right)} A\left(\frac{x}{\sigma}\right)\ dx}\right] \\ & = 0. \end{aligned} \tag{C.34}$$

We can proceed similarly for the Fisher Information in the case of direct detection. We recall the expression:

$$\mathcal{I}_{\text{Direct}}(\theta) = \int_{-\infty}^{\infty} \frac{I'(x)^2}{I(x)} dx, \text{ with } I'(x,\theta) = \frac{\partial I(x,\theta)}{\partial \theta} \tag{C.35}$$

In the case of two point sources, the normalized intensity profile can be written with the previous ASF as : $I(x) = (|A(x+\theta)|^2 + |A(x-\theta)|^2)/2$. One can write for its first derivative:

$$\forall x \in \mathbb{R},\ \lim_{\theta \to 0} I'(x,\theta) = \lim_{\theta \to 0} \operatorname{Re}\left[\overline{A'(x+\theta)}A(x+\theta) - \overline{A'(x-\theta)}A(x-\theta)\right] = 0 \tag{C.36}$$

$$\Rightarrow \lim_{\theta \to 0}\ \mathcal{I}_{\text{Direct}}(\theta) = 0. \tag{C.37}$$

Proof 5 : Fisher Information Inequality On Aggregated Measurements. Let m_S be an aggregated measurement over a collection S of measurement functions from orthogonal modes, *i.e.* $m_S(\theta) = \sum_{q \in S} m_q(\theta)$, all corrupted by Poisson noise. At a location θ where $\forall q,\ m_q(\theta) > 0$, we have the following measurement function and Fisher Information:

$$\mathcal{I}_S(\theta) = \frac{\left(\sum_{q \in S} m'_q(\theta)\right)^2}{\sum_{q \in S} m_q(\theta)} \tag{C.38}$$

We can simplify the notations for the current location into sets (m_q) and (m'_q) and observe that:

$$\left(\sum_{q \in S} m_q\right)\left(\sum_{q \in S} \frac{m'^2_q}{m_q}\right) - \left(\sum_{q \in S} m'_q\right)^2 = \sum_{q \in S} \sum_{r \in S, r > q} \frac{m_r}{m_q} m'^2_q + \frac{m_q}{m_r} m'^2_r - 2m'_q m'_r \tag{C.39}$$

Here, if m'_q or m'_r are equal to zero or their product is negative, and with the previous positivity constraint, the sum is clearly positive. Otherwise, one can write:

$$\frac{1}{m'_q m'_r}\left(\frac{m_r}{m_q} m'^2_q + \frac{m_q}{m_r} m'^2_r - 2m'_q m'_r\right) = \frac{m_r}{m_q}\frac{m'_q}{m'_r} + \frac{m_q}{m_r}\frac{m'_r}{m'_q} - 2 \geq 0 \tag{C.40}$$

As we have, $m_r m'_q / m_q m'_r > 0$ and $\forall x > 0,\ x + 1/x \geq 2$. Thus, (C.39) is always positive and we can conclude with the inequality:

$$\frac{\left(\sum_{q\in S} m_q'(\theta)\right)^2}{\sum_{q\in S} m_q(\theta)} = \mathcal{I}_S(\theta) \leq \sum_{q\in S} \frac{m_q'(\theta)^2}{m_q(\theta)} = \sum_{q\in S} \mathcal{I}_q(\theta) \tag{C.41}$$

References

[1] H. Miller and J. W. Strange, "The electrical reproduction of images by the photoconductive effect," Proceedings of the Physical Society, Volume 50, Number 3, pp. 374–384 (1937)

[2] H. Iams and A. Rose, "Television Pickup Tubes with Cathode-Ray Beam Scanning," in Proceedings of the Institute of Radio Engineers, vol. 25, no. 8, pp. 1048–1070, Aug. 1937. doi: 10.1109/JRPROC.1937.228423.

[3] E. G. Schoultz, "Procédé et appareillage pour la transmission des images mobiles à distance". Patent No. FR 539,613. Office National de la Propriété industrielle. Filed in 1921, patented in 1922.

[4] W. S. Boyle, G. E. Smith, "Charge Coupled Semiconductor Devices", Bell System Technical Journal Vol. 49, no. 4, pp. 587–593, (1970).

[5] G. E. Moore, "Cramming more components onto integrated circuits," Electronics Magazine 38(8), (1965).

[6] H. H. Barrett, "Objective assessment of image quality: effects of quantum noise and object variability," J. Opt. Soc. Am. A 7, 1266-1278 (1990).

[7] H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, "Objective assessment of image quality. II. Fisher information, Fourier crosstalk, and figures of merit for task performance," J. Opt. Soc. Am. A 12, 834-852 (1995).

[8] H. H. Barrett, C. K. Abbey, and E. Clarkson, “Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions,” J. Opt. Soc. Am. A 15, 1520-1535 (1998).

[9] E. R. Dowski and W. T. Cathey, “Extended depth of field through wave-front coding” Applied Optics, **34**, 1859-1866 (1995).

[10] K. Strehl, “Aplanatische und fehlerhafte Abbildung im Fernrohr”, Zeitschrift für Instrumentenkunde 15 (Oct. 1895), 362-370.

[11] K. Strehl, “Über Luftschlieren und Zonenfehler”, Zeitschrift für Instrumentenkunde, 22 (July 1902), 213-217.

[12] A. Maréchal, “Étude des effets combinés de la diffraction et des aberrations géométriques sur l’image d’un point lumineux”. Rev. Opt. 2: 257277 (1947).

[13] V. Mahajan, “Strehl ratio for primary aberrations in terms of their aberration variance,” J. Opt. Soc. Am. 73, 860-861 (1983).

[14] B. Liege, R. Tessieres, F. Guichard, E. Knauer, H. Nguyen, Dxo Labs, “Optical system provided with a device for augmenting its depth of field”, U.S. Patent No. 8331030 (2008).

[15] Norbert Wiener, “Extrapolation, Interpolation, and Smoothing of Stationary Time Series”, Wiley, (1949).

[16] S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, “Optimization by Simulated Annealing” Science, 220(**4598**), 671-680 (1983).

[17] F. Guichard, H. Nguyen, R. Tessières, M. Pyanet, I. Tarchouna, F. Cao, “Extended depth-of-field using sharpness transport across color channels”, Proc. SPIE 7250, Digital Photography V, 72500N (January 19, 2009); doi:10.1117/12.805915.

[18] S. Bagheri, P. Silveira, and D. de Farias, “Analytical optimal solution of the extension of the depth of field using cubic-phase wavefront coding. Part I. Reduced-complexity

approximate representation of the modulation transfer function", J. Opt. Soc. Am. A 25, 1051-1063 (2008).

[19] S. Bagheri, P. Silveira, and G. Barbastathis, "Signal-to-noise-ratio limit to the depth-of-field extension for imaging systems with an arbitrary pupil function", J. Opt. Soc. Am. A 26, 895-908 (2009).

[20] T. Vettenburg, N. Bustin, and A. Harvey, "Fidelity optimization for aberration-tolerant hybrid imaging systems", Opt. Express 18, 9220-9228 (2010).

[21] Y. Takahashi and S. Komatsu, "Optimized free-form phase mask for extension of depth of field in wavefront-coded imaging", Opt. Lett. 33, 1515-1517 (2008).

[22] Victor P. Pauca, Robert J. Plemmons, Sudhakar Prasad, Todd C. Torgersen, Joseph van der Gracht, "Integrated optical-digital approaches for enhancing image restoration and focus invariance", Proc. SPIE 5205, Advanced Signal Processing Algorithms, Architectures, and Implementations XIII, 348 (December 31, 2003), doi:10.1117/12.505309.

[23] G. Carles, A. Carnicer, S. Bosch, "Iterative design of mesh-defined phase masks for wavefront coding", Proc. SPIE 7723, Optics, Photonics, and Digital Technologies for Multimedia Applications, 77231O (6 May 2010); doi: 10.1117/12.854374

[24] W. Chi and N. George, "Computational imaging with the logarithmic asphere: theory", J. Opt. Soc. Am. A 20, 2260-2273 (2003).

[25] F. Diaz, F. Goudail, B. Loiseaux, and J. Huignard, "Comparison of depth-of-focus-enhancing pupil masks based on a signal-to-noise-ratio criterion after deconvolution", J. Opt. Soc. Am. A 27, 2123-2131 (2010).

[26] Qingguo Yang, Liren Liu, Jianfeng Sun, Optimized phase pupil masks for extended depth of field, Optics Communications, Volume 272, Issue 1, 1 April 2007, Pages 56-66, ISSN 0030-4018, http://dx.doi.org/10.1016/j.optcom.2006.11.021.

[27] D. Reshidko and J. Sasian, "Current trends in miniature camera lens technology", SPIE Newsroom. DOI: 10.1117/2.1201602.006327, 19 February 2016.

[28] D. Donoho, “Compressed sensing”, in *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, April 2006.

[29] M. Neifeld and P. Shankar, “Feature-specific imaging”, Appl. Opt. 42, 3379-3389 (2003).

[30] M. A. Neifeld and J. Ke, “Optical architectures for compressive imaging”, in *Appl. Opt. 46*, 5293-5303 (2007)

[31] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly and R. Baraniuk, “A New Compressive Imaging Camera Architecture using Optical-Domain Compression”, PCI IV at SPIE EI, CA, Jan. 2006

[32] E. J. Candes and T. Tao, “Decoding by linear programming,” in IEEE Transactions on Information Theory, vol. 51, no. 12, pp. 4203-4215, Dec. 2005. doi: 10.1109/TIT.2005.858979

[33] E. J. Candès, J. K. Romberg, and T. Tao, “Stable Signal Recovery from Incomplete and Inaccurate Measurements”, in *Communications on Pure and Applied Mathematics*, Vol. LIX, 12071223 (2006).

[34] E. J. Candès and J. Romberg, “ℓ1-magic: Recovery of Sparse Signals via Convex Programming”, January 2005.

[35] J. M. Bioucas-Dias and M. A. T. Figueiredo, “A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration”, in *IEEE Trans. on Image Proc.*, vol. 16, iss. 12, December 2007.

[36] J. A. Tropp and A. C. Gilbert, “Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit”, in *IEEE Trans. on Inform. Theory*, vol. 53, iss. 12, December 2007.

[37] Y. Guoshen, G. Sapiro, S. Mallat, “Solving Inverse Problems With Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity”, in *IEEE TSP*, vol. 21, no. 5, pp.2481-2499 1057-7149 (2012)

[38] A. Ashok, L.-C. Huang, and M. A. Neifeld, “Information optimal compressive sensing: static measurement design”, in *JOSA A* 30, 831-853 (2013)

[39] A. Mahalanobis, “Recent Results of Infra-red Compressive Sensing”, in *Classical Optics 2014*, OSA Technical Digest (online) (Optical Society of America, 2014), paper CM2D.1.

[40] R. Kerviche, N. Zhu, and A. Ashok, “Information-optimal Scalable Compressive Imaging System”, in *Classical Optics 2014*, OSA Technical Digest (online) (Optical Society of America, 2014), paper CM2D.2.

[41] R. Kerviche, N. Zhu and A. Ashok, “Information optimal scalable compressive imager demonstrator”. In 2014 IEEE International Conference on Image Processing, ICIP 2014. (pp. 2177-2179). [7025439] Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/ICIP.2014.7025439

[42] J. Wang, M. Gupta and A. C. Sankaranarayanan, “LiSens- A Scalable Architecture for Video Compressive Sensing”, Computational Photography (ICCP), 2015 IEEE International Conference on, Houston,TX, 2015, pp. 1-9. doi: 10.1109/ICCPHOT.2015.7168369

[43] H. Chen, M. S. Asif, A. C. Sankaranarayanan and A. Veeraraghavan, “FPA-CS: Focal plane array-based compressive imaging in short-wave infrared”, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2358-2366. doi: 10.1109/CVPR.2015.7298849

[44] J. P. Dumas, M. A. Lodhi, W. U. Bajwa, and M. C. Pierce, “Computational imaging with a highly parallel image-plane-coded architecture: challenges and solutions”, in *Opt. Express 24*, 6145-6155 (2016)

[45] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, Amit Ashok; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 449-458

[46] L. Huang, M. Neifeld, and A. Ashok, ”Face recognition with non-greedy information-optimal adaptive compressive imaging,” Appl. Opt. 55, 9744-9755 (2016).

[47] W. B. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space”, Contemporary Mathematics, 26:189206, 1984.

[48] W. B. Johnson, J. Lindenstrauss and G. Schechtman, “Extensions of lipschitz maps into Banach spaces”, I. J. Math. (1986) 54: 129. doi:10.1007/BF02764938.

[49] L. M. Novak, G.J. Owirka and C.M. Netishen, “Radar target identification using spatial matched filters,” Pattern Recognition, 27(4)(1994), 607617.

[50] K. Palaniappan, R. M. Rao, G. Seetharaman, “Wide-Area Persistent Airborne Video: Architecture and Challenges,” pp 349-371 in “Distributed Video Sensor Networks”, doi:10.1007/978-0-85729-127-1_24, Springer London (2011).

[51] E. Blasch, G. Seetharaman, S. Suddarth, K. Palaniappan, G. Chen, H. Ling and A. Basharat, “Summary of methods in Wide-Area Motion Imagery (WAMI),” Proc. SPIE 9089, Geospatial InfoFusion and Video Analytics IV; and Motion Imagery for ISR and Situational Awareness II, 90890C (June 19, 2014); doi:10.1117/12.2052894.

[52] C. E. Shannon., “A Mathematical Theory of Communication”, Bell Syst. Tech. J., 27:379-423, 623-656, July-October 1948

[53] R. M. Fano, “Transmission of Information: A Statistical Theory of Communication”, MIT: New York, 1961.

[54] V. A. Kovalevskij, “The problem of character recognition from the point of view of mathematical statistics”, Character Readers and Pattern Recognition, pp. 3-30, Spartan: New York, 1968. Russian edition 1965.

[55] A. Rényi, “On measures of information and entropy,” Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960. pp. 547561.

[56] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions.”, Bull. Calcutta Math. Soc. 35, (1943). 99109.

[57] M. A. Neifeld, A. Ashok, and P. K. Baheti, “Task-specific information for imaging system analysis,” J. Opt. Soc. Am. A 24, B25-B41 (2007).

[58] P. Baheti and M. Neifeld, "Adaptive feature-specific imaging: a face recognition example," Appl. Opt. 47, B21-B31 (2008).

[59] A. Ashok, P. K. Baheti, and M. A. Neifeld, "Compressive imaging system design using task-specific information," Appl. Opt. 47, 4457-4471 (2008).

[60] P. K. Baheti and M. A. Neifeld, "Recognition using information-optimal adaptive feature-specific imaging," J. Opt. Soc. Am. A 26, 1055-1070 (2009).

[61] D. Erdogmus and J.C. Principe, "Lower and Upper Bounds for Misclassification Probability Based on Renyi's Information," The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology (2004) 37: 305. doi:10.1023/B:VLSI.0000027493.48841.39

[62] J. C. Principe, "Information Theoretic Learning, Renyi's Entropy and Kernel Perspectives", Springer 2010.

[63] K. Kampa, E. Hasanbelliu and J. C. Principe, "Closed-form cauchy-schwarz PDF divergence for mixture of Gaussians," The 2011 International Joint Conference on Neural Networks, San Jose, CA, 2011, pp. 2578-2585.

[64] H.G. Hoang, B.N. Vo, B.T. Vo and R. Mahler, "The Cauchy-Schwarz divergence for poisson point processes," 2014 IEEE Workshop on Statistical Signal Processing (SSP), Gold Coast, VIC, 2014, pp. 240-243.

[65] C. Hegde, A. C. Sankaranarayanan, W. Yin and R. G. Baraniuk, "NuMax: A Convex Approach for Learning Near-Isometric Linear Embeddings," in IEEE Transactions on Signal Processing, vol. 63, no. 22, pp. 6109-6121, Nov.15, 2015.

[66] Y. Li, C. Hegde, A. C. Sankaranarayanan, R. Baraniuk and K. F. Kelly, "Compressive image acquisition and classification via secant projections", Journal of Optics, Volume 17.6, 2015.

[67] M. Dunlop-Gray, P. Poon, D. Golish, E. Vera, and M. Gehm, “Experimental demonstration of an adaptive architecture for direct spectral imaging classification,” Opt. Express 24, 18307-18321 (2016).

[68] L. Huang, M. Neifeld, and A. Ashok, “Face recognition with non-greedy information-optimal adaptive compressive imaging,” Appl. Opt. 55, 9744-9755 (2016).

[69] A. Mahalanobis, R. Muise and S. Roy, ”Efficient target detection using an adaptive compressive imager,” in IEEE Transactions on Aerospace and Electronic Systems, vol. 50, no. 4, pp. 2528-2540, October 2014.

[70] J. Huang and A. Ashok, “Information Optimal Compressive X-ray Threat Detection,” in Imaging and Applied Optics 2015, OSA Technical Digest (online) (Optical Society of America, 2015), paper CTh2E.4.

[71] L. Bottou, “Large-Scale Machine Learning with Stochastic Gradient Descentm,” in: Lechevallier Y., Saporta G. (eds) Proceedings of COMPSTAT’2010. Physica-Verlag HD (2010).

[72] M. Zinkevich, Ma. Weimer, L. Lihong and A. J. Smola, “Advances in Neural Information Processing Systems 23,” in Advances in Neural Information Processing Systems 23, pp. 2595-2603, Curran Associates, Inc. (2010).

[73] Lord Rayleigh F.R.S., “Investigations in optics, with special reference to the spectroscope”, Philosophical Magazine Series 5, **8**, 49 (1879).

[74] R.A. Fisher, “Theory of Statistical Estimation”, Mathematical Proceedings of the Cambridge Philosophical Society, 22(5), pp. 700725. doi: 10.1017/S0305004100009580 (1925).

[75] C.R. Rao, “Information and the accuracy attainable in the estimation of statistical parameters”. Bulletin of the Calcutta Mathematical Society. 37: 81–89. MR 0015748 (1945).

[76] H. Cramer, “Mathematical Methods of Statistics”. Princeton, NJ: Princeton Univ. Press. ISBN 0-691-08004-6. OCLC 185436716 (1946).

[77] J. A. Nelder and R. Mead, “A Simplex Method for Function Minimization,” Comput J 1965; 7 (4): 308-313. doi: 10.1093/comjnl/7.4.308

[78] M. Abramowitz, I. Stegun, “Handbook of Mathematical Functions,” U.S. National Bureau of Standards, Washington, DC, (1964).

[79] S. Hell and J. Wichmann, “Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy,” Opt. Lett. 19, 780-782 (1994).

[80] T. Klar and S. Hell, “Subdiffraction resolution in far-field fluorescence microscopy,” Opt. Lett. 24, 954-956 (1999).

[81] S.T. Hess, T.P.K. Girirajan and M.D. Mason, “Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy,” Biophys J. 2006 Dec 1; 91(11): 42584272. doi: 10.1529/biophysj.106.091116 (2006).

[82] C. Helstrom, “Resolution of point sources of light as analyzed by quantum detection theory,” in IEEE Transactions on Information Theory, vol. 19, no. 4, pp. 389-398, Jul 1973. doi: 10.1109/TIT.1973.1055052.

[83] M. Tsang, R. Nair, and X.M. Lu, “Quantum Theory of Superresolution for Two Incoherent Optical Point Sources,” Phys. Rev. X **6**, 031033 (2016).

[84] R. Nair and M. Tsang, “Far-Field Superresolution of Thermal Electromagnetic Sources at the Quantum Limit,” Phys. Rev. Lett. **117**, 190801 (2016).

[85] C. Lupo and S. Pirandola, “Ultimate Precision Bound of Quantum and Subwavelength Imaging,” Phys. Rev. Lett. **117**, 190802 (2016).

[86] H. Krovi, S. Guha, J.H. Shapiro, “Attaining the quantum limit of passive imaging,” arXiv:1609.00684 (2016).

[87] R. Nair and M. Tsang, "Interferometric superlocalization of two incoherent optical point sources," Opt. Express 24, 3684-3701 (2016).

[88] Z.S. Tang, K. Durak and A. Ling, "Fault-tolerant and finite-error localization for point emitters within the diffraction limit," Opt. Express 24, 22004-22012 (2016).

[89] W.K. Tham, H. Ferretti and A.M. Steinberg, "Beating Rayleighs Curse by Imaging Using Phase Information," Phys. Rev. Lett. 118, 070801 (2017).