IMAGE PROCESSING FOR QUANTA IMAGE SENSORS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Omar A. Elgendy

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF THESIS APPROVAL

Dr. Stanley H. Chan, Chair

School of Electrical and Computer Engineering

Dr. Charles A. Bouman

School of Electrical and Computer Engineering

Dr. Edward J. Delp

School of Electrical and Computer Engineering

Dr. Xiaojun Lin

School of Electrical and Computer Engineering

Approved by:

Dr. Dimitrios Peroulis

Head of the School of Electrical and Computer Engineering

This thesis is dedicated to my beloved wife, my parents
and my children. I could not write it without their support.

ACKNOWLEDGMENTS

"Praise be to God for guiding us to this. We would have never been guided if God had not guided us.", Quran 7:43.

I would like to express my deepest appreciation and gratitude to my doctoral advisor Professor Stanley H. Chan. I consider myself as a lucky person to work under his supervision. I have learned a lot from him, both in scientific and personal aspects. Actually, this work would not have been developed in this form without his guidance, scientific support, and fruitful discussions.

I would like to thank Professor Charles A. Bouman, Professor Edward J. Delp and Professor Xiaojun Lin for their invaluable criticism and fruitful feedback for my dissertation. I would like to show my appreciation to Professor Charles A. Bouman for teaching me ECE637 and ECE641 courses. These courses had a great impact on my understanding of image processing science in theoretical and practical aspects.

I sincerely acknowledge my colleagues in Statistical Signal and Image Processing Lab (SSIP) lab: Abhiram Gnanasambandam, Chengxi Li, Joon Hee Choi, Nicholas Chimitt, Xiangyu Qu, Xiran Wang and Zhiyuan Mao for their continuous encouragement, support and sharing experiences and knowledge.

In addition, for their faithful friendship and encouragement, it is pleasure to thank my friends Abobakr Alabbasi, Ahmed Atef, Amr Mohamed, Amr Ebeid, Camilo Aguilar, Mostafa Abdalla, Ashraf Youssef.

Finally, I would like to express my profound gratitude to my lovely wife whose patience, continuous encouragement, support, and constant presence enabled me to complete this work. I would like to thank my parents and sisters for their support and encouragement to me.

I believe that PhD is not the end of my lifelong learning journey. It is the beginning of my pursuit of knowledge. *"My Lord! Increase me in knowledge.", Quran 20:114*

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Elgendy, Omar A. PhD, Purdue University, August 2019. Image Processing for Quanta Image Sensors. Major Professor: Stanley H. Chan.

Since the birth of charge coupled devices (CCD) and the complementary metal-oxide-semiconductor (CMOS) active pixel sensors, pixel pitch of digital image sensors has been continuously shrinking to meet the resolution and size requirements of the cameras. However, shrinking pixels reduces the maximum number of photons a sensor can hold, a phenomenon broadly known as the full-well capacity limit. The drop in full-well capacity causes drop in signal-to-noise ratio and dynamic range.

The Quanta Image Sensor (QIS) is a class of solid-state image sensors proposed by Eric Fossum in 2005 as a potential solution for the limited full-well capacity problem. QIS is envisioned to be the next generation image sensor after CCD and CMOS since it enables sub-diffraction-limit pixels without the inherited problems of pixel shrinking. Equipped with a massive number of detectors that have single-photon sensitivity, the sensor counts the incoming photons and triggers a binary response "1" if the photon count exceeds a threshold, or "0" otherwise. To acquire an image, the sensor oversamples the space and time to generate a sequence of binary bit maps. Because of this binary sensing mechanism, the full-well capacity, signal-to-noise ratio and the dynamic range can all be improved using an appropriate image reconstruction algorithm. The contribution of this thesis is to address three image processing problems in QIS: 1) Image reconstruction, 2) Threshold design and 3) Color filter array design.

Part 1 of the thesis focuses on reconstructing the latent grayscale image from the QIS binary measurements. Image reconstruction is a necessary step for QIS because the raw binary measurements are not images. Previous methods in the literature

use iterative algorithms which are computationally expensive. By modeling the QIS binary measurements as quantized Poisson random variables, a new non-iterative image reconstruction method based on the Transform-Denoise framework is proposed. Experimental results show that the new method produces better quality images while requiring less computing time.

Part 2 of the thesis considers the threshold design problem of a QIS. A spatially-varying threshold can significantly improve the reconstruction quality and the dynamic range. However, no known method of how to achieve this can be found in the literature. The theoretical analysis of this part shows that the optimal threshold should match with the underlying pixel intensity. In addition, the analysis proves the existence of a set of thresholds around the optimal threshold that give asymptotically unbiased reconstructions. The asymptotic unbiasedness has a phase transition behavior. A new threshold update scheme based on this idea is proposed. Experimentally, the new method can provide good estimates of the thresholds with less computing budget compared to existing methods.

Part 3 of the thesis extends QIS capabilities to color imaging by studying how a color filter array should be designed. Because of the small pixel pitch of QIS, crosstalk between neighboring pixels is inevitable and should be considered when designing the color filter arrays. However, optimizing the light efficiency while suppressing aliasing and crosstalk in a color filter array are conflicting tasks. A new optimization framework is proposed to solve the problem. The new framework unifies several mainstream design criteria while offering generality and flexibility. Extensive experimental comparisons demonstrate the effectiveness of the framework.

1. INTRODUCTION

Miniaturization has become the main theme in CCD and CMOS image sensor industry recently. There are two complementary motives behind miniaturing the pixel size. First, pixel miniaturization improves the spatial resolution by allowing more pixels for the same sensor size. High spatial resolution is necessary for obtaining good image quality, especially in low-light scenarios, and it avoids common artifacts resulting from scene undersampling. Second, pixel miniaturization allows smaller sensor size for the same spatial resolution. Sensors with small sensor sizes are particularly useful for smart phones cameras which have space limitations. As a result, restless effort has been exerted in image sensor industry to reduce pixel sizes to dimensions even less than diffraction limit of light. This is facilitated by the continuous improvement in CMOS fabrication technologies in semiconductor foundries. However, reducing the pixel size results in decreasing the amount of charge that it can hold before saturating, which is known formally as the pixel's full-well capacity (FWC). Reducing FWC, in turn, leads to a drop in signal-to-noise (SNR) ratio and a drop in dynamic range. These problems have been a fundamental impediment against pixel miniaturization, and efficient solutions are still required to overcome it, both on the hardware and signal processing sides.

This dissertation studies a new type of sensors, called the Quanta Image sensor (QIS), which is proposed as a potential hardware solution for the previously mentioned miniaturization problems. It tackles three signal processing problems which are essential for the success of QIS hardware solution: *Image Reconstruction*, *Threshold Design* and *Color Filter Design*. This thesis provides solutions for the first two problems that enhance monochrome image reconstruction for QIS. It also presents solutions for the third problem that improves and facilitates color image reconstruction for QIS.

The outline of this introductory chapter is as follows. First, it presents in Section 1.1 a quick discussion on the history of photography. Then, it motivates the QIS solution in Section 1.2 and shows how it evolved over time both in the hardware side and the signal processing side. Afterwards, it shows in Section 1.3 the motivations for tackling the image reconstruction, threshold design, and color filter design problems. Finally, it presents in Section 1.4 the thesis outline and a summary of contributions.

1.1 Photography: From Camera Obscura to Computational Photography

In this section, we will give a quick summary for the history of photography starting from the camera obscura and ending with the state-of-the-art computational photography.

1.1.1 Camera Obscura

The ancestor of modern day camera is the *Camera Obscura*, a Latin name that means "dark chamber" or the *Pinhole Camera*. As the name suggests, it comprises a small dark room with light entering to it through a tiny hole or "aperture" and reflecting on the opposite wall to show an inverted image of the scene outside (See Figure 1.1). This idea was discussed as early as the 5th and 4th centuries B.C. by the Chinese philosopher Mo Ti [1], and the Greek mathematicians Aristotle and Euclid [2], respectively. However, the first conceptual analysis and experimental realization were published by the Muslim scientist Alhazen [3] in his book: *Book of Optics* written in Cairo during the early 11th century.

The quality of the projected image depends on the pinhole size. Too large pinhole generates a bright, but blurry image due to geometrical blur, and too small pinhole leads to a dim blurry image due to diffraction blur. Even at the optimal pinhole size, the projected image is still dim because of the small pinhole size. Later on, lenses were deployed to alleviate this trade-off by enlarging the aperture size to absorb more light while focusing this light to produce sharper images.

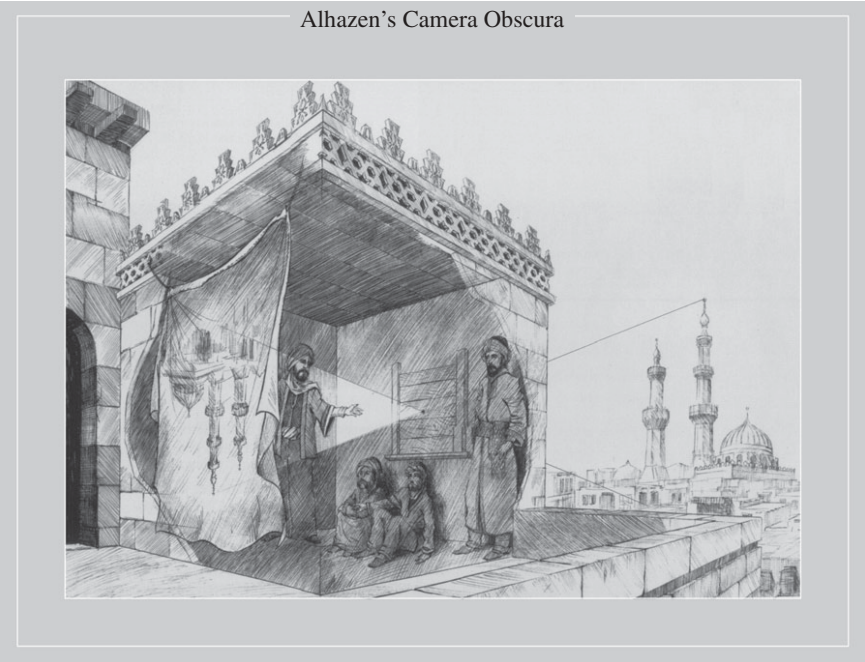


Fig. 1.1. A schematic for the conception of the *Camera Obscura* by Alhazen in his book *Book of Optics* written in Cairo between years 1011 and 1021. Adapted from [3]

Another important question is: How to make use of the projected image? For a long time ago, people were using it to observe Sun eclipses without harming the eyes, and by the 16th century, artists were using it for drawing objects with lots of details by tracing the projected image on a drawing paper. However, there was a strong need for saving the projected image. This need was satisfied by invention of chemical photography.

1.1.2 Chemical Photography

The concept of using chemical compounds to save an image was conceived in 1727 when Schulze discovered that silver nitrate salt is darkened when exposed to sunlight [4]. This inspired Nièce, a French inventor, in 1816 to use a paper coated with silver chloride salt to capture images [4, 5]. However, this approach could not store

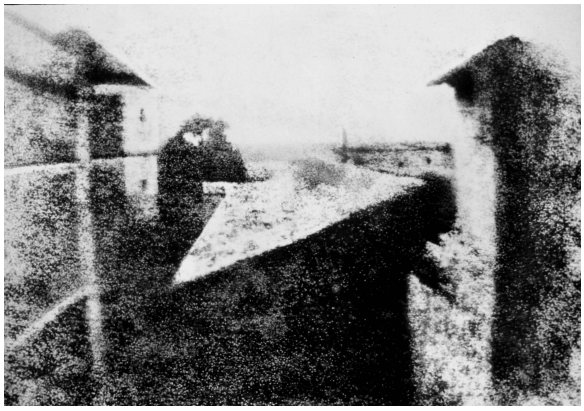


Fig. 1.2. The first permanent photograph captured by Nièce in 1826 at Saint-Loup-de-Varennes, France [6]

the image permanently. After some trials, he managed to capture the first permanent image (See Figure 1.2) in 1826 with 8 hours of exposure. It was not until 1839 when Daguerre presented the first stable photographic process, namely the daguerreotype process [5], which was commercialized after that. One remaining challenge was to replicate the images. Talbot solved this problem in 1841 by using papers coated with silver iodide [5] and Archer improved it in 1851 by allowing multiple copies from a single negative. However, the negatives required immediate development in no more than 10 minutes. Maddox solved this problem in 1861 allowing, for the first time, hand-held cameras.

Modern photography began when Eastman presented the first transparent photographic film and film roll as a replacement for the photographic plate in 1885 and 1889, respectively. This film comprises a light sensitive material placed on paper, which is transferred on glass after exposure, and then printed. With some modifications, Oskar Barnack presented the 35mm film in 1925 [7] which became the standard film for a long time after that.

On another front, color imaging was first introduced in 1861 when Sutton captured the first color image using a method proposed by J. C. Maxwell. Sutton captured three images with red, green and blue filters, then projected them on a screen using

the same filters to add up giving a color image. Afterwards, the Lumière brothers invented in 1906 the first practical color photography plate by using a mosaic of three color filters mounted on a glass layer which is placed under the light sensitive layer. However, it required longer exposure time because of the decreased light sensitivity after putting the glass layer. Mannes and Godowsky, Jr. presented in 1935 the first popular color film: the Kodachrome. However, it could only be processed in Kodak labs since its processing was too complex for commercial users.

1.1.3 Digital Photography

The era of digital photography started with the conception of the photoelectric effect: a phenomenon that was first observed by Hertz in 1887, and characterized by Einstein in 1905 [8] [1]. This finding is significant for photography because it shows that light falling on a matter can alter it properties in a way proportional to the light intensity. This led to the invention of photodetector that replaces photographic film by using photodiodes instead of chemical compounds to save the image. This is advantageous for two main reasons: 1) By resetting the photodetector, we do not need to replace it after each capture as we do with photographic film, and 2) By digitization, we have more flexibility to process the captured image.

The first practical realization of this technology was done by at AT&T Bell Labs by Boyle and Smith in 1968 when they invented an imaging semiconductor circuit: The *Charge-Coupled Device (CCD)* [9] [2]. A CCD comprises an array of photodetectors and shift register that works as a conveyor belt. After light exposure, every photodetector accumulates an electric charge proportional to the light intensity falling onto it, then the shift register transfer the charges to feed it to a charge amplifier sequentially to be converted into voltages. On a digital camera, these voltages, which constitutes the captured image, are digitized and stored in memory.

[1] Einstein won the Nobel prize in physics for this work in 1921

[2] Boyle and Smith won the Nobel prize in physics for this work in 2009

Another breakthrough in digital photography occured when Eric Fossum and his team in NASA's Jet Propulsion Laboratory (JPL) invented the CMOS active pixel sensor (APS). It quickly became a ubiquitous imaging technology in mobile imaging for its lower power consumption and smaller size compared to CCD. Also, its compatibility with the standard CMOS fabrication technology enabled it to benefit from Moore's scaling law by continuously shrinking the pixel size for resolution enhancement. In addition, several on-chip functionalities are added such as Analog-to-Digital conversion (ADC) and Image signal processors (ISPs) to improve the image quality. For these reasons, there is a consensus that CMOS image sensor is the second generation of digital image sensors after CCD.

To enable color imaging on CCD and CMOS image sensors, the most popular technique is to place a color filter array on top of the sensor so that each pixel gets color information of the falling light. Color filters are organized in a certain way so that the captured image is a mosaic pattern of different colors. In 1976, Bayer proposed the Bayer pattern for color filter arrays. This pattern is a periodic replica of a 2×2 color kernel that comprises 1 red, 2 green and 1 blue color filters, where the green proportion is more than red and blue proportion because the eye is more sensitive to light in the green color subband. A color image is reconstructed from the mosaicked image by a process called *demosaicking*. The demosaicking process in its basic form for is an interpolation process that aims at reconstructing the two missing colors at each pixel.

1.1.4 Computational Photography

In spite of the continuous development in cameras from the camera obscura to CMOS image sensors, the main idea was similar: light entering through an aperture and focused on a detector by a lens to form an image. In other words, the traditional camera performs a passive and conservative sampling of incoming light without any further processing. With the current advances in fabrication technology, the camera

can be equipped with more computational power to form a *computational camera*. Instead of passively capturing photons, this additional intelligence enables the camera to *compute* pictures instead of sensing them [10].

Computational photography has opened the door for numerous ideas and applications that take advantage of on-chip computations. We will present some representative ideas for brevity.

- **High Dynamic Range Imaging** [11,12]: An image with high dynamic range (HDR) is acquired by combining multiple images with low dynamic range (LDR) having different exposures. The fusion weights are computed post-capture to yield a high dynamic range with lots of details in both dark and bright regions of the image.

- **Multi-Aperture Imaging** [13]: In contrast to conventional imaging, a point in the scene is mapped to multiple points on the sensor in multi-aperture imaging by slightly shifting the image sensor away from the focal plane and using micro-lenses. This leads to multiple sub-images of the scene. An image is reconstructed by warping the sub-images and combining them. There are two main benefits behind this architecture. First, it helps in capturing the depth information by measuring how certain features are located within the sub-images. Second, it offers a new method for color imaging by replacing the per-pixel color filter array with *per-aperture* color filter array. This method is more robust to crosstalk as it is restricted to neighboring pixels having the same color. However, it loses spatial resolution because every point in the scene is sampled by three apertures having the three color filters.

- **Light Field Imaging** [14]: In contrast to conventional image which captures the intensity of light at each pixel, light field imaging aims at capturing the intensity and the direction of light. This generalizes the image from a 2D projection of the scene to 4D projection. Using an idea similar to multi-aperture

imaging and some post-capture computations, this concept can be realized using the conventional CMOS image sensor.

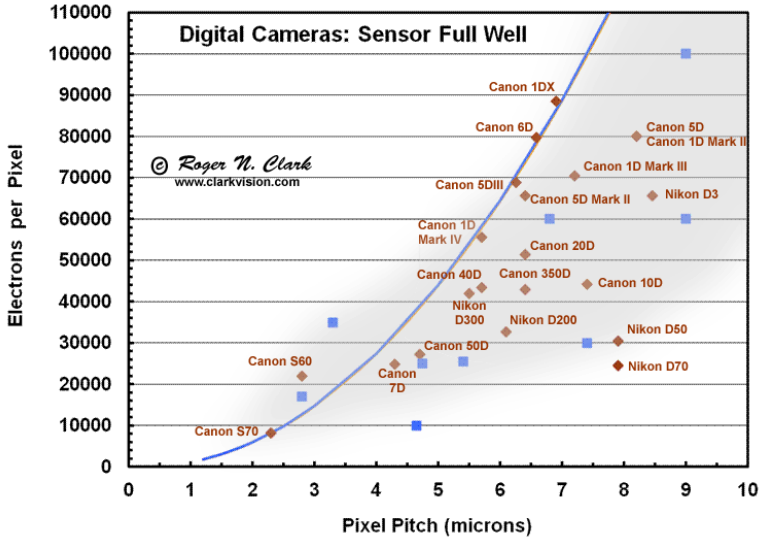
- **Compressive Sensing Imaging** [15]: The idea of compressive sensing is to reconstruct a sparse signal from multiple linear measurements obtained by projecting the signal using multiple random linear projectors. This enables sampling the scene at sub-Nyquist sampling rates. By leveraging the compressive sensing concept, a scene can be reconstructed by using different random masks that obtain random linear measurements of the scene. This reduces the required resolution of the sensor to even a single pixel.

1.2 Quanta Image Sensor

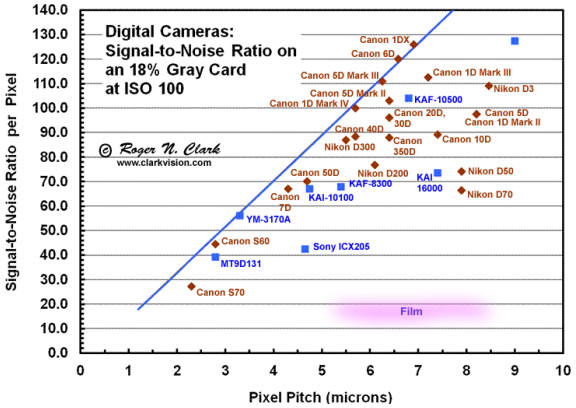
Quanta Image Sensor (QIS) is a class of solid-state image sensors designed to solve the miniaturization problems of CMOS sensor and envisioned to be the next generation imaging device after it. Originally proposed by Eric Fossum in 2005 [16], the sensor has gained significant momentum in the past decade, both in terms of hardware design [17–19] and image processing [20–24].

1.2.1 Motivation

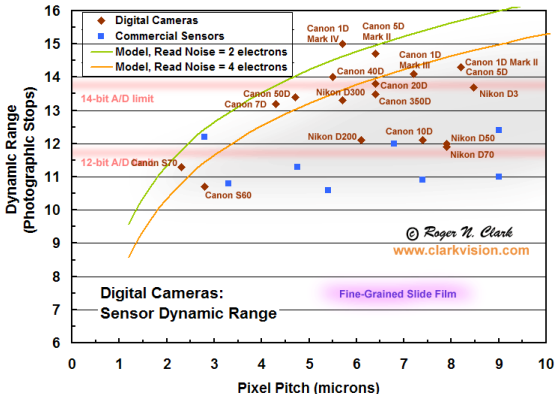
The main trend in image sensor industry is *Miniaturization*. This trend aims at shrinking the pixel size to improve the sensor resolution for increased image detail, or to decrease the camera size at the same resolution for increased flexibility. This trend is shown in the curves of Figure 1.3 [25] which are collected from the specifications of different cameras during the past years. From the curves in Figure 1.3, we notice that as the pixel pitch is decreased, the full-well capacity is reduced (Figure 1.3(a).) This, in turn, causes a drop in signal-to-noise ratio (Figure 1.3(b)) and a drop in dynamic range (Figure 1.3(c).) Using the current image sensor technology, these fundamental



(a) Full-Well Capacity vs. Pixel Pitch [25]



(b) SNR vs. Pixel Pitch [25]



(c) Dynamic Range vs. Pixel Pitch [25]

Fig. 1.3. As we decrease the pixel pitch, or alternatively the pixel size, the (a) Full-Well capacity decreases, and this results in a decrease in (b) SNR, and (c) Dynamic range.

problems are inevitable, and they require sophisticated algorithms to reduce their effect.

QIS aims at solving these problems by providing a new paradigm in imaging. The main idea is to allow the pixel size to decrease as much as possible (e.g. 100−200 nm pitch [26]) to form miniature pixels, called *jots*, with intentionally low FWC (1−200 photoelectrons [26]). Each jot has sub-electron readout noise (i.e., readout noise with

standard deviation less than 0.3 electron [27, 28]) which enables it to have single-photon sensitivity and photon counting capability. The jot counts every incoming photon and produces a binary response "1" if the photon count exceeds a threshold q, and "0" otherwise. By making $q <$ FWC, the resulting signal has high SNR because of its binary nature, and this solves the first miniaturization problem of poor SNR.

Definitely, the binary quantization of photon counts leads to significant distortion in the output signal. To compensate for this aggressive quantization of light, QIS oversamples the light signal in space and time by having huge spatial resolution (e.g., 10^9 pixels per sensor with 200nm pitch per jot [28]) and huge temporal resolution or frame rate (e.g., 100k fps as reported in [29]), respectively. As a result, each output gray-scale pixel is formed by locally processing a 3d spatial-temporal kernel or a "cubicle" of $K \times K \times T$ binary jots, where K is the spatial kernel size and T is the number of temporal frames. This processing is usually referred to as *binning* and it is frequently used in low-light image processing to mitigate noise. By efficient processing of the cubicle of jots, the output pixel represents the incoming light intensity on these jots. Figure 1.4 shows the QIS image formation process. The high spatial-temporal oversampling of QIS increases its dynamic range to levels even higher than CMOS and CCD, and this solves the second miniaturization problem of low dynamic range.

Another useful property of QIS is its programmability or flexibility. For a fixed cubicle volume K^2T, the cubicle shape can be varied according to the scene allowing for a spatial-temporal resolution trade-off. For example, a cubicle of $K\sqrt{T} \times K\sqrt{T} \times 1$ jots can be used for ultra-fast applications when the resolution is not so critical. Alternatively, a cubicle of $1 \times 1 \times K^2T$ jots can be used to obtain high-resolution images for static scenes. These two scenarios are depicted in Figure 1.5. The cubicle shape can be adjusted post-acquisition according to the scene properties. This adjustment can be temporally-varying with frames, or spatially-varying within one frame, or both.

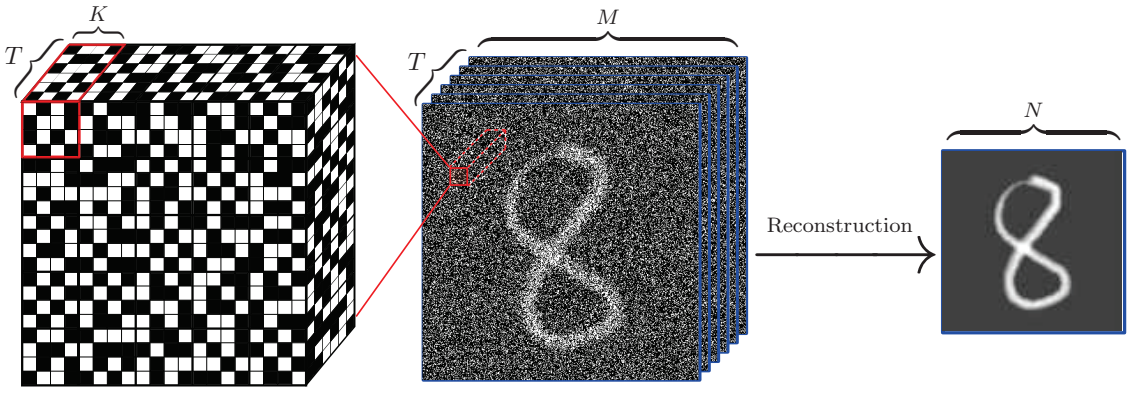
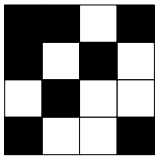
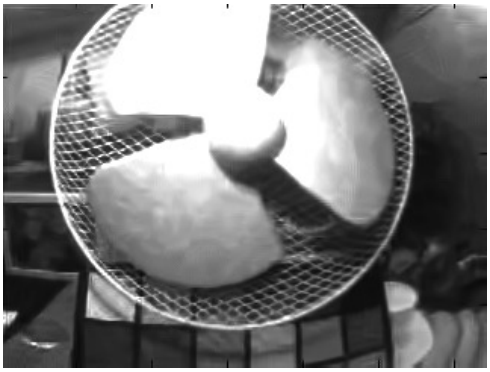


Fig. 1.4. Image reconstruction of QIS data. Given T binary bit planes having high resolution $M \times M$, the reconstruction algorithm processes each $K \times K \times T$ cubicle of jots to form the $N \times N$ gray-scale image shown on the right, where $N = M/K$.



(a) $4 \times 4 \times 1$ reconstruction kernel

(b) $1 \times 1 \times 16$ reconstruction kernel



(c) Moving Fan Image [30]

(d) Static High Resolution Image

Fig. 1.5. To improve temporal resolution, a cubicle of $4 \times 4 \times 1$ jots (a) can be used for ultra-fast applications when the spatial resolution can be small like image (c). Alternatively, to improve spatial resolution, a cubicle of $1 \times 1 \times 16$ jots (b) can be used to obtain high-resolution images for static scenes like image (d)

1.2.2 Evolution of QIS Concept

QIS belongs to the family of photon-counting devices. These photon-counting devices have been known for a long time. Some better-known examples are the electron-multiplying charge-coupled device (EMCCD) [31, 32], single-photon avalanche diode (SPAD) [29, 30, 33], Geiger-mode avalanche photodiode (GMAPD) [34], etc. These sensors have reached a mature level in their design and fabrication; however, their applications are limited to scientific and military purposes. On the other hand, QIS is designed to compete in the commercial market beside its scientific and military applications.

The concept of QIS was first proposed by Fossum in 2005 as a solution for sub-diffraction limit pixels. The sensor was called the digital film sensor, and later the quanta image sensor [35–37]. After the introduction of QIS, researchers in EPFL developed a similar concept called the Gigavision camera [21, 38, 39], where they mainly tackled the image reconstruction problem assuming the presence of suitable hardware. Recently, teams at the University of Edingburgh [30, 33, 40] and EPFL [41, 42] have made new progresses in QIS using binary single-photon detectors. In industry, Rambus Inc. (Sunnyvale, CA) has developed binary image sensors for high dynamic range imaging [43–45]. Table 1.1 lists several recent QIS prototypes that are available or are currently being developed. As a comparison we also show a Canon 5D Mark III CMOS camera. Among many different features, the most noticeable is the frame rate. For example, SPS SPAD can be operated at 20k fps. SwissSPAD can even achieve 156k fps. Both are significantly faster than a standard CMOS camera.

Recently, a startup company [47] has been established to develop and realize practical prototypes of QIS. Resolution is expected to rise from 1024×1024 in [48] to 10240×10240, and total power per bit is expected to be reduced from 16pJ/bit to 9.9pJ/bit as mentioned in the conference presentation of [48].

Beside alleviating the miniaturization problems, the single-photon sensitivity of QIS nominates it as a perfect candidate for low-light applications such as astronomy

Table 1.1.
List of QIS Prototypes and Parameters

Camera	Canon 5D CMOS	EMCCD [46]	GMAPD [34]	SPC SPAD [33]	SwissSPAD [29]	Fossum QIS [37]
Price	$\$5,000$	$\$20,000$	Prototype	Prototype	Prototype	Prototype
Resolution	4096×2160	1024×1024	256×256	320×240	512×128	1376×768
Pixel Pitch (μm)	6.25	13	25	8	24	3.6
Full-well Capacity	69 ke-	80 ke-	-	$56 - 125$ e-	-	$1 - 250$ e-
Frame Rate (fps)	6	$26 - 92$	8×10^3	2×10^4	1.56×10^5	1×10^3
Sensor data rate	88.6 Mbps	0.48 Gbps	0.52 Gbps	1.54 Gbps	10.24 Gbps	1 Gbps

[49], night-vision [50], and medical imaging [51–53]. Also, its huge frame rate allows it to track ultra-fast objects in low-light with high resolution [54]. Specifically, the high frame rates of QIS simplify the tracking of fast moving objects because the local shift in consecutive frames is limited and can be easily estimated. QIS has also been used for nuclear engineering [55], depth and reflectivity reconstruction [56], and recently in quantum random number generation used in cryptography [41, 57].

1.3 Motivation

In this section, we show our motivation to study the following image processing problems for QIS: 1) QIS image reconstruction (Section 1.3.1), 2) QIS threshold design (Section 1.3.2) and 3) Color filter design for QIS (Section 1.3.3)

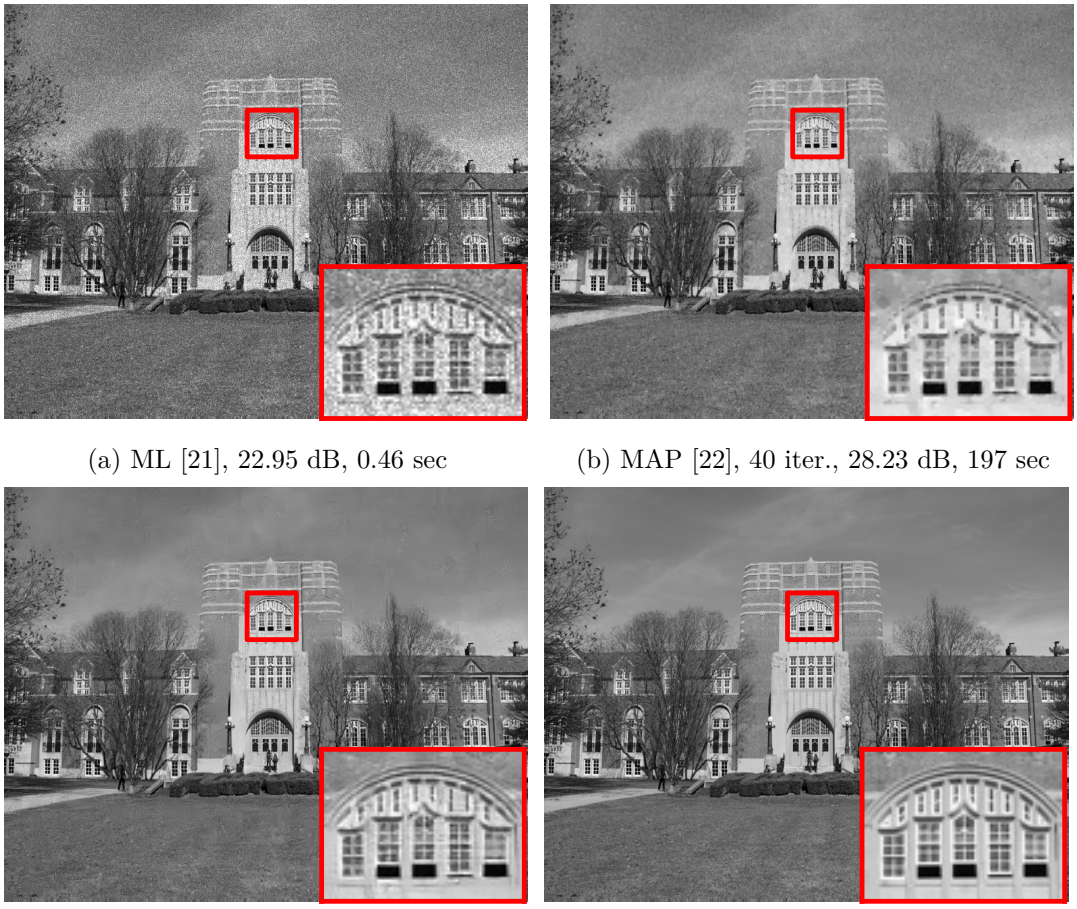
1.3.1 QIS Image Reconstruction

To obtain a grayscale image from QIS binary measurements, an image reconstruction algorithm is required. This algorithm should be extremely fast in order to handle the high frame rates of QIS. In addition, it should have the flexibility to reconstruct images with spatially invariant or spatially varying threshold.

A simple way to reconstruct a grayscale image from the binary frames is *digital integration*. Each output gray-scale pixel is formed by simply averaging bits in each $K \times K \times T$ cubicle. The quantization threshold q can be fixed for all time frames [35, 38, 40, 58] or it can be a temporal sequence of decreasing or increasing thresholds for dynamic range improvement [43–45]. However, this simple averaging approach requires T to be large enough to have a practical dynamic range. This wastes the temporal oversampling of QIS. A smart integration technique is proposed in [59,60] as a solution where the frames are summed in overlapping temporal windows. However, this overlapping introduces colored noise in the output.

Another approach is to formulate image reconstruction as an inverse problem [61], and use *statistical estimation* techniques to solve it. Maximum likelihood estimation (MLE) criterion is used in [20, 21, 39, 62–64], yet the results are noisy because the problem is ill-conditioned. To produce clean results, Maximum-A-Posterior (MAP) criterion is used with different priors such as sparsity-based priors [65–67] and total-variation prior [22]. Except for some simplified assumptions where the MLE problem gives a closed-form solution [21], iterative techniques are used to solve the inverse problem such as dynamic programming [65], interior point algorithms [20], gradient descent [21], simplex search [62], random walks [64], ADMM [22], and unrolled ISTA iterations implemented by a neural network [66, 67].

On one hand, iterative techniques used to get the MAP estimate are not suitable for ultra-fast imaging tasks. On the other hand, the fast MLE closed-form solution is too noisy. Hence, in order for QIS to be a practical competitive for CCD and CMOS, a fast and efficient image reconstruction algorithm is required. Figure 1.6 shows reconstructed images by ML criterion [21], MAP criterion in [22], and our proposed method compared to ground truth. Our proposed method can achieve the best of two worlds: It can reconstruct a clean image like MAP estimate in short time like ML estimate.



(a) ML [21], 22.95 dB, 0.46 sec

(b) MAP [22], 40 iter., 28.23 dB, 197 sec

(c) Our method, 29.50 dB, 2.33 sec

(d) Ground Truth

Fig. 1.6. Simulated QIS data and the reconstructed gray-scale images using different reconstruction methods. The results show that our method reconstructs high quality image in short time. In this experiment, we spatially oversample each pixel by $K = 4 \times 4$ binary bits and we use $T = 5$ independent temporal measurements. Quantization threshold is fixed to $q = 1$ in all methods.

1.3.2 QIS Threshold Design

Optimal threshold design for QIS is important as it directly affects the dynamic range of an image. Figure 1.7 illustrates an example where we simulate the raw binary data acquired by a QIS using a uniform threshold q. When q is low, most of the bits in the raw input are "1". The reconstructed image is therefore an over-exposed image.

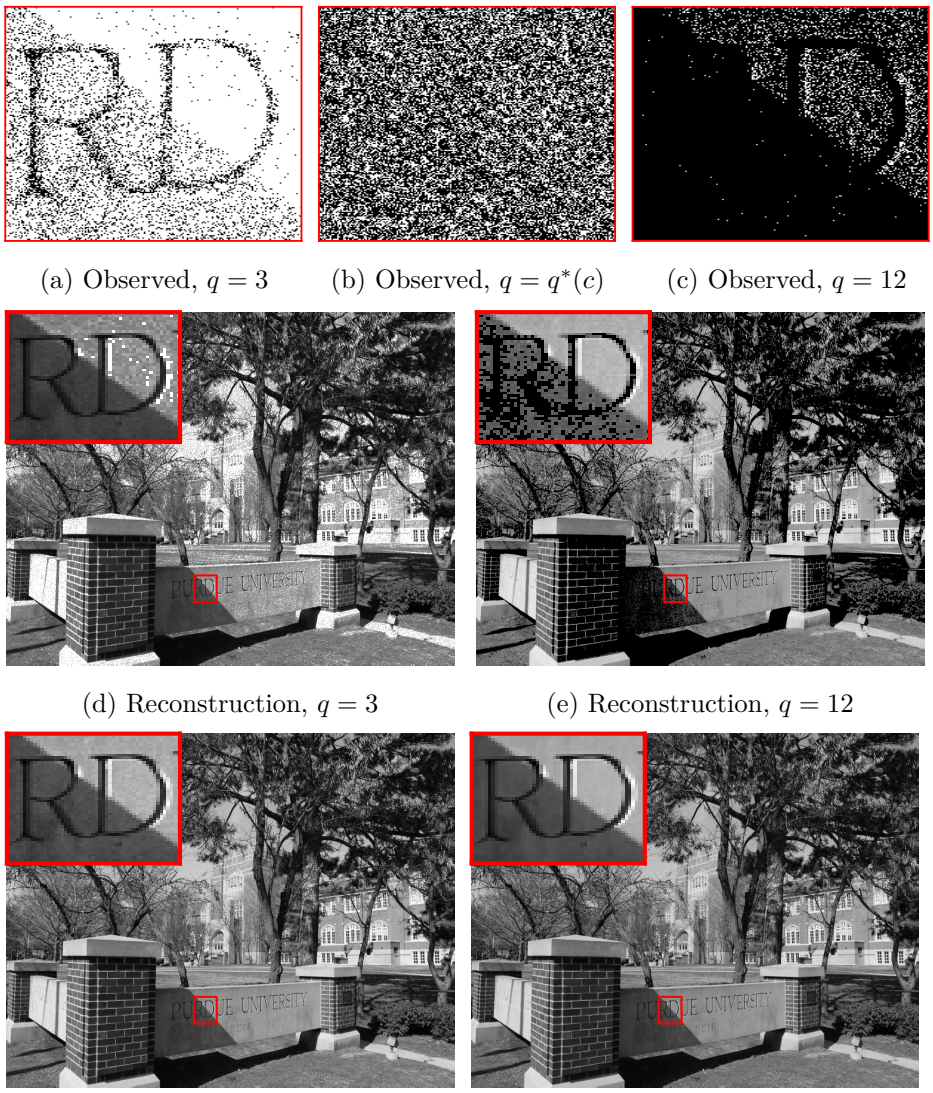
On the other hand, when q is high, most of the bits in the raw input are "0". The reconstructed image is then under-exposed. In both cases, it is evident from the simulation that a uniform threshold has limited performance. A better way is to allow q to vary spatially so that a pixel (or a group of pixels) has its own threshold value. The result in Figure 1.7(d) shows the reconstruction result using a spatially varying threshold obtained from our proposed technique, which is clearly better than the uniform thresholds.

Existing work on QIS threshold design study can be summarized into three classes of methods.

- Markov Chain [62]. The Markov Chain method developed by Hu and Lu [62] is a time-sequential update scheme. A Markov Chain probability is used to control how the threshold q of each jot should be increased or decreased in every frame. While the method has provable convergence, the threshold of each jot has to be updated sequentially in time. In contrast, our proposed method allows a group of jots to share the same threshold. As a result, our proposed method has significantly faster rate of convergence.

- Conditional Reset [43–45]. The conditional reset method is a hardware solution proposed by Vogelsang and colleagues. The idea is to take a sequence of frames with ascending (or descending) uniform thresholds, and digitally integrate the sequence to form a gray-scale image. The drawback of the method, besides the additional hardware cost of the per-pixel reset transistors, is the limited quality of the reconstructed image. For the same number of frames, our proposed method produces better images.

- Checkerboard Threshold [6]. This method constructs a checkerboard of thresholds by alternating two threshold values q_1 and q_2. The optimality criterion of q_1 and q_2 is based on minimizing the Cramér-Rao lower bound (CRLB) integrated over a range of light intensities, which is essentially an average case result. Our proposed method obtains the optimal threshold for each pixel. This per-pixel



(a) Observed, $q = 3$ (b) Observed, $q = q^*(c)$ (c) Observed, $q = 12$

(d) Reconstruction, $q = 3$ (e) Reconstruction, $q = 12$

(d) Reconstruction, $q = q^*(c)$ (e) Ground Truth

Fig. 1.7. Simulated QIS data and the reconstructed gray-scale images using different thresholds. Top row: The binary measurements obtained using thresholds $q = 3$, $q = q^*(c)$, and $q = 12$. Bottom figures: The maximum likelihood estimates obtained from the binary measurements, with comparison to the ground truth. The results show that our spatially varying threshold $q^*(c)$ offers the best reconstruction. In this experiment, we spatially oversample each pixel by $K = 2 \times 2$ binary bits and we use $T = 25$ independent temporal measurements.

optimization has higher reconstruction performance compared to checkerboard threshold.

1.3.3 Color Filter Arrays Design

Despite the rapid advancement in QIS hardware [28, 35, 68] and algorithms [21, 23, 24, 69], all reported findings, to-date, are based on monochromatic data. The first color QIS imaging is only recently proposed by Gnanasambandam et al. [70], where they demonstrated how to reconstruct a color image from the sensor with a Bayer color filter array. In this thesis, we discuss how to design color filter array for better image acquisition.

A color filter array (CFA) is a mask placed on top of the sensor to select (filter) wavelengths. As light passes through the color filter array, the resulting image is a mosaic pattern of the three tri-stimulus RGB colors. Traditionally, CFA is organized as a periodic replica of a 2D kernel called the *color atom*. The de-facto color atom used in the industry is the Bayer pattern [71] because of its simplicity and the readily available demosaicking algorithms include [72–81]. More sophisticated CFAs have been proposed [82–92] to improve the Bayer CFA.

When designing a CFA, there are three factors that should be taken into consideration:

- **Aliasing**: Since color filtering is a sampling process, aliasing happens when the sampling rate is less than Nyquist. Aliasing causes false color artifacts at color edges, called the *Moirè* artifacts [83]. Color filters that are susceptible to aliasing, such as the Bayer CFA, require sophisticated demosaicking algorithms to suppress the Moirè artifacts. In contrast, a robust CFA can use simple demosaicking algorithms.

- **Sensitivity**: Since CFA is a filter, it blocks part of the incoming light. This reduces the sensor sensitivity and makes the image more susceptible to noise. A good CFA design should maximize the sensitivity by allowing transparent or "panchromatic" color filters that block as few wavelengths as possible.

- **Crosstalk**: Crosstalk can be either optical or electrical [93]. If not treated, crosstalk will make colors look pale or de-saturated. Crosstalk desaturation is

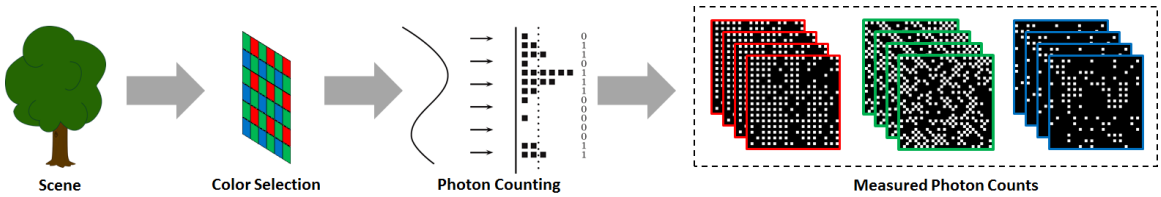


Fig. 1.8. **QIS Imaging Model**. When the scene image arrives at the sensor, the CFA first selects the wavelength according to the colors. Each color pixel is then sensed using a photon-detector and reports a binary value based on whether the photon counts exceeds certain threshold or not. The measured data contains three subsampled sequences, each representing a measurement in the color channel.

corrected by pixel-wise multiplication of the RGB color vector using a color correction matrix. However, color correction enhances residual noise in the image [93, 94]. The situation is worsen in QIS because of its small size.

The three factors above are conflicting: Optimizing one generally degrades the others. For conventional CMOS image sensors, crosstalk is not severe, and so most CFA designs in the literature consider aliasing and sensitivity only. The only available work on QIS color filter array design is by Anzagira and Fossum [93]. However, aliasing was not adequately handled.

The design framework we propose in this thesis is a unification of several mainstream CMOS-based color filter arrays. To put our work in the proper context in the literature, we here list a few of the better known results.

- **Spatial CFA Design**: By suppressing the Moirè artifacts and crosstalk while keeping the demosaicing algorithm simple, Lukac and Plataniotis [82] proposed a CFA and compared it with other CFAs using a universal demosaicking method. However, their work did not provide a mathematical framework to analyze the CFA optimality.

- **Spatio-Spectral CFA Design**: Hirakawa and Wolfe [83] proposed a method through the spatial and spectral domain analysis. Their CFA reduces aliasing

in the frequency domain, and possesses high sensitivity and numerical stability. Condat [95] extended the framework by optimizing luminance and chrominance sensitivity. He defined a new form of orthogonality between chrominance channels in frequency domain. Hao et al. [86] and Wang et al. [87] proposed a framework based on symbolic discrete Fourier transform (DFT). Their CFA maximizes the numerical stability of linear demosaicking process under aliasing and physical constraints.

- **Learning-based CFA Design**: By minimizing the average error on a color dataset, Lu and Vetterli [84] used an iterative algorithm to solve a least squares CFA design problem. Chakrabarti [96] and Henz et al. [97] proposed to learn the optimal CFA pattern by using a deep neural network.

Besides these mainstream CFA design frameworks, there are a number of other CFA designs such as [85,88–92]. On the hardware side, [98] and [99] took into account that color filter fabrication technology lags the image sensor technology in terms of miniaturization. They proposed a hardware-friendly CFA assuming the color filter size is 1.5× pixel size.

1.4 Thesis Outlines and Contributions

The goal of this thesis is three-fold. First, it proposes an efficient and fast QIS image reconstruction algorithm. This algorithm should have the flexibility to handle spatially-varying threshold, which is the best option according to Section 1.3.2. Second, it presents an optimal threshold design methodology and provide theoretical justifications for it. Finally, it presents an optimization framework for CFA design that encompasses aliasing, sensitivity and crosstalk in a unified model. This is the first work that incorporates a quantitative crosstalk metric in an optimization framework for CFA design.

As for QIS image reconstruction, our contributions are summarized as follows.

- First, we extend the ADMM algorithm proposed in [22] to spatially-varying threshold. As mentioned in Section 1.3.1, threshold has a critical effect of the reconstruction quality, where a poorly selected threshold will result in either an under-exposed image or an over-exposed image. However, most algorithms can only handle spatially invariant threshold.

- Second, we propose a non-iterative algorithm for reconstructing clean QIS images in short time. This algorithm is based on a *Transform-Denoise* framework. Under certain conditions, the ML solution has a closed-form expression which requires summing the bits in each cubicle. By observing the distribution of the summed bits, we can use a suitable variance stabilizing transform to make the noise spatially-invariance. Hence, we can use any standard image denoising algorithm to remove this noise before applying the ML expression. Experimental results shows the effectiveness of our method in terms of quality and speed compared to other methods.

As for QIS threshold design, we have two major contributions:

- First, we provide a rigorous theoretical analysis of the performance limits of QIS image reconstruction as a function of the threshold. These results form the basis of our subsequent discussions of the threshold update scheme. Some results are known, e.g., the signal-to-noise ratio is a function of the Fisher Information [6, 64], but a number of new results are shown. In particular, we show that (i) the maximum likelihood estimate has a closed-form expression in terms of the incomplete Gamma function, (ii) the oracle threshold can be derived in closed-form by maximizing the signal-to-noise ratio, and (iii) the image reconstruction has a phase transition behavior.

- Second, we propose an efficient threshold update scheme based on our theoretical results. The new scheme is a bisection method which iteratively updates the threshold *without* the need of reconstructing the image. By checking whether the proportion of one's and zero's approaches 0.5 in a spatial-temporal cubicle,

the threshold is guaranteed to be near optimal. Compared to other existing threshold update schemes such as [62] and [43–45], the new scheme offers significantly faster rate of convergence. We also demonstrate how the dynamic range can be extended for high dynamic range (HDR) imaging.

As for Color filter array design, the main contribution is a general and flexible framework for CFA design. Compared to the existing CFA design framework, the new framework is able to simultaneously (Section 5.2)

- Improve CFA's luminance and chrominance sensitivity,
- Reduce aliasing between luminance and chrominance channels,
- Suppress crosstalk between spectral sub-bands, and
- Enforce orthogonality between chrominance channels to permit simple linear demosaicking.

The design framework is presented in the form of optimization. We have two designs: A convex optimization and a non-convex optimization. In addition to the formulation, we also present an algorithm to solve the non-convex optimization. (Section 5.3)

For performance evaluation of different CFAs on QIS images, we propose in Section 5.4 a universal demosaicking pipeline. This pipeline comprises a demosaicking by frequency selection algorithm for removing the CFA masking effect followed by a color correction step for removing the desaturation effect of crosstalk. Experimental evaluation on the Kodak and McMaster color datasets shows the robustness of our proposed CFAs compared to other CFAs in literature.

The work in this thesis has resulted in the following publications:

- **O. A. Elgendy**, and S. H. Chan, "Color Filter Arrays for Quanta Image Sensors," submitted to *IEEE Transactions on Computational Imaging*, June 2019.
- A. Gnanasambandam, **O. A. Elgendy**, J. Ma, and S. H. Chan, "Megapixel Photon-Counting Color Imaging using Quanta Image Sensor," *Optics Express*,

vol. 27, no. 12, pp. 17298-17310, June 2019. [Online]. Available: https://www.osapublishing.org/oe/abstract.cfm?uri=oe-27-12-17298

- **O. A. Elgendy**, and S. H. Chan, "Optimal Threshold Design for Quanta Image Sensor," *IEEE Transactions on Computational Imaging*, vol. 4, no. 1, pp. 99-111, March 2018.

- S. H. Chan, **O. A. Elgendy**, and X. Wang, "Images from Bits: Non-iterative Image Reconstruction for Quanta Image Sensors," *MDPI Sensors* Special Issue on Photon-Counting Image Sensors, vol. 16, no. 11, November 2016, Article number: 1961. [Online]. Available: https://www.mdpi.com/1424-8220/16/11/1961

- **O. A. Elgendy**, and S. H. Chan, "Image reconstruction and threshold design for quanta image sensors," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP'16)*, Phoenix, AZ, USA, 2016, pp. 978–982.

In parallel to the work done in this thesis, we worked on the following publications:

- J. H. Choi, **O. A. Elgendy**, and S. H. Chan, "Optimal Combination of Image Denoisers," *IEEE Transactions on Image Processing*, Early Access, March 2019.

- J. H. Choi, **O. A. Elgendy**, and S. H. Chan, "Image Reconstruction for Quanta Image Sensors using Deep Neural Networks," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*, Calgary, AB, 2018, pp. 6543–6547

- S. H. Chan, X. Wang, and **O. A. Elgendy**, "Plug-and-Play ADMM for Image Restoration: Fixed Point Convergence and Applications," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 84-98, March 2017.

2. QIS IMAGING MODEL

In this chapter, we provide an overview of the QIS imaging model. The model has been previously discussed in several papers, e.g., [21–24]. Readers interested in details can refer to these papers for further explanations. Without loss of generality, we assume 1-dimensional signals where extension to 2-dimensional signals is straightforward.

2.1 Spatial Oversampling

We denote the discrete version of the light intensity as a vector $\boldsymbol{c} = [c_0, \ldots, c_{N-1}]^T$, where $n = 0, \ldots, N-1$ specify the spatial coordinates. We assume that c_n is normalized to the range $[0, 1]$ for all n so that there is no scaling ambiguity. To model the actual light intensity, we multiply c_n by a constant α to yield αc_n, where $\alpha > 0$ is a fixed scalar constant.

The continuous version of the light intensity field $\lambda(x)$ is obtained by convolving with a non-negative interpolation kernel $\phi(x)$ as follows

$$\lambda(x) = \frac{N}{\tau} \sum_{n=0}^{N-1} c_n \phi(Nx - n), \tag{2.1}$$

where τ is the exposure time. Examples of the interpolation kernel include

- Box-car kernel

$$\beta(x) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } 0 \leq x \leq 1. \\ 0, & \text{if otherwise} \end{cases} \tag{2.2}$$

- Cardinal B-splines

$$\beta_k(x) = \left(\underbrace{\beta * \ldots * \beta}_{k+1 \text{ times}} \right) (x + \frac{k}{2}) \tag{2.3}$$

As a sampling device, QIS uses $M \gg N$ jots to sample the light field intensity $\lambda(x)$. The ratio $K \stackrel{\text{def}}{=} M/N$ is known as the spatial oversampling factor. Assume that the mth jot covers the interval $[\frac{m}{M}, \frac{m+1}{M}] \subset [0,1]$ for $m \in \{0, \ldots, M-1\}$. Denote by θ_m, the total light exposure integrated in the mth jot during exposure time period $[0, \tau]$. Hence, we can calculate θ_m as follows.

$$\begin{aligned}\theta_m &\stackrel{\text{def}}{=} \alpha \int_0^\tau \int_{m/M}^{(m+1)/M} \lambda(x)\, dx\, dt \\ &= \alpha\tau \langle \lambda(x), \beta(Mx - m) \rangle \end{aligned} \tag{2.4}$$

where $\beta(x)$ is the box function defined in X, and $\langle ., . \rangle$ represents the standard L^2 inner product between two continuous functions f and g, which is defined as $\langle f, g \rangle = \int_{-\infty}^{\infty} f(x)g(x)\, dx$. Substituting (2.1) in (2.4) , we get

$$\begin{aligned}\theta_m &= \alpha\tau \left\langle \frac{N}{\tau} \sum_{n=0}^{N-1} c_n \phi(Nx - n), \beta(Mx - m) \right\rangle \\ &= \alpha \sum_{n=0}^{N-1} c_n \left\langle N\phi(Nx - n), \beta(Mx - m) \right\rangle \\ &\stackrel{(a)}{=} \alpha \sum_{n=0}^{N-1} c_n \left\langle \phi(x), \beta\left(\frac{M(x+n)}{N} - m\right) \right\rangle \\ &= \alpha \sum_{n=0}^{N-1} c_n \left\langle \phi(x), \beta\left(Kx - (m - Kn)\right) \right\rangle \\ &\stackrel{(b)}{=} \alpha \sum_{n=0}^{N-1} c_n g_{m-Kn} \end{aligned} \tag{2.5}$$

where step (a) is obtained by using the change of variables $Nx - n \to x$, and step (b) is obtained by defining the discrete filter

$$g_k \stackrel{\text{def}}{=} \langle \phi(x), \beta\left(Kx - k\right) \rangle . \tag{2.6}$$

In multi-rate signal processing notation, (2.5) represents K-fold upsampling of a signal c_n followed by filtering with discrete low-pass filter $\{g_k\}$. In matrix notation, (2.5) can be rewritten as follows

$$\boldsymbol{\theta} = \alpha \boldsymbol{G} \boldsymbol{c}, \tag{2.7}$$

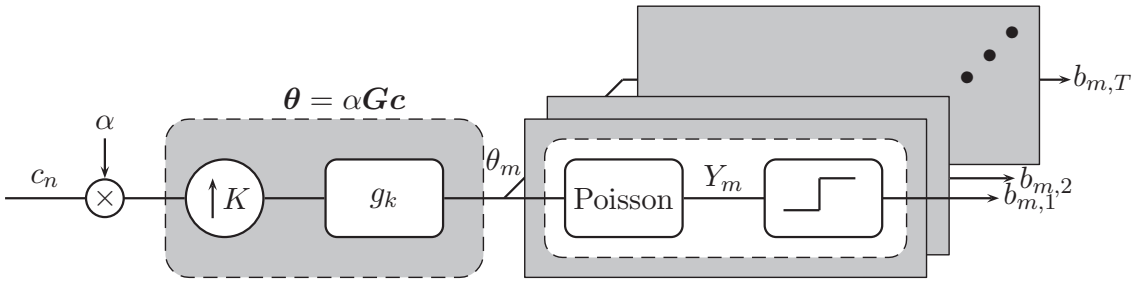


Fig. 2.1. Block diagram illustrating the image formation process of QIS.

where $\boldsymbol{\theta} \in \mathbb{R}^M$ is the light exposure vector, $\boldsymbol{c} \in \mathbb{R}^N$ is the light intensity vector, and $\boldsymbol{G} \in R^{M \times N}$ is a circulant matrix representing the upsampling-filtering process. The overall process is depicted in the right block in Figure 2.1.

When $\phi(x) = \beta(x)$, i.e., the interpolation filter has box-car kernel, the filter $\{g_k\}$ can be calculated as follows.

$$g_k \stackrel{\text{def}}{=} \langle \beta(x), \beta\left(Kx - k\right) \rangle \tag{2.8}$$

$$= \int_{-\infty}^{\infty} \beta(x) \beta\left(Kx - k\right) dx \tag{2.9}$$

$$= \begin{cases} \frac{1}{K}, & \text{if } k \in \{0, 1, \ldots, K-1\} \\ 0, & \text{if otherwise} \end{cases} \tag{2.10}$$

which means that $\{g_k\}$ has a box-car kernel that sum to one. In this case, the matrix $\boldsymbol{G}$ can be defined as

$$\boldsymbol{G} = \frac{1}{K} \boldsymbol{I}_{N \times N} \otimes \mathbf{1}_{K \times 1}, \tag{2.11}$$

where $\mathbf{1}_{K \times 1}$ is a vector of all ones and $\otimes$ denotes the Kronecker product. The box-car kernel assumption is typically reasonable, because on each QIS jot there is a micro-lens to focus the incident light. Although previous papers, e.g., [21, 22], do not make such assumption, in this thesis we decide to use the simplified $\boldsymbol{G}$, for otherwise the theoretical analysis will become very complicated. Nevertheless, in the supplementary material we show comparison between a general $\boldsymbol{G}$ and the simplified $\boldsymbol{G}$. The performance gap is usually insignificant.

2.2 Truncated Poisson Process

We assume that the operating speed of QIS is significantly faster than the scene motion. Therefore, for a given scene $\boldsymbol{c}$ (and also $\boldsymbol{\theta}$), we are able to acquire a set of T independent measurements. We illustrate this using the T channels in Figure 2.1.

Photons impinge on the m-th jot during the t-th independent measurement according to a Poisson process, with mean value equal to the light exposure θ_m on this jot, i.e., the photon count $Y_{m,t}$ follows the Poisson distribution:

$$\mathbb{P}(Y_{m,t} = y_{m,t}) = \frac{\theta_m^{y_{m,t}} e^{-\theta_m}}{y_{m,t}!}, \tag{2.12}$$

where $m = 0, \ldots, M-1$ denotes the m-th jot of the QIS and $t = 0, 1, \ldots, T-1$ denotes the t-th independent measurement in time. Denoting $q \in \mathbb{N}$ as the quantization threshold, the final observed binary measurement $B_{m,t}$ is a truncation of $Y_{m,t}$:

$$B_{m,t} = \begin{cases} 0, & \text{if } Y_{m,t} < q. \\ 1, & \text{if } Y_{m,t} \geq q \end{cases}$$

The probability mass function of $B_{m,t}$ is given by

$$\mathbb{P}(B_{m,t} = b_{m,t}) = \begin{cases} \sum_{k=0}^{q-1} \frac{\theta_m^k e^{-\theta_m}}{k!}, & \text{if } b_{m,t} = 0, \\ \sum_{k=q}^{\infty} \frac{\theta_m^k e^{-\theta_m}}{k!}, & \text{if } b_{m,t} = 1. \end{cases} \tag{2.13}$$

The goal of image reconstruction is to recover the underlying image $\boldsymbol{c}$ from the binary measurements $\mathcal{B} = \{B_{m,t} \mid m = 0, \ldots, M-1, \text{and } t = 0, \ldots, T-1\}$. A pictorial illustration of the reconstruction is shown in Figure 1.4.

2.3 Properties of Truncated Poisson Processes

The probability mass function of $B_{m,t}$ in (2.13) is Bernoulli. However, the right hand side of (2.13) involves infinite sums which are difficult to interpret. To simplify

the equations, we consider the upper incomplete Gamma function $\Psi_q : \mathbb{R}^+ \to [0, 1]$ defined in [100] as:

$$\Psi_q(\theta) \stackrel{\text{def}}{=} \frac{1}{\Gamma(q)} \int_\theta^\infty t^{q-1} e^{-t} dt, \quad \text{for } \theta > 0,\ q \in \mathbb{N}.$$

where $\Gamma(q) = (q-1)!$ is the standard Gamma function. The incomplete Gamma function allows us to rewrite the infinite sums in (2.13) using the following identity [100]:

$$\Psi_q(\theta) = \sum_{k=0}^{q-1} \frac{\theta^k}{k!} e^{-\theta}. \tag{2.14}$$

Consequently, the probabilities in (2.13) become

$$\begin{aligned} \mathbb{P}(B_{m,t} = 0) &= \Psi_q(\theta_m), \\ \mathbb{P}(B_{m,t} = 1) &= 1 - \Psi_q(\theta_m). \end{aligned} \tag{2.15}$$

Example 1 *In the special case of $q = 1$, we obtain:*

$$\mathbb{P}(B_{m,t} = 0) = \frac{1}{\Gamma(1)} \int_{\theta_m}^\infty t^0 e^{-t} dt = e^{-\theta_m},$$

which coincides with the results shown in [21] and [22].

The incomplete Gamma function $\Psi_q(\theta)$ is a decreasing function of θ because the first order derivative of $\Psi_q(\theta)$ with respect to θ is negative:

$$\frac{d}{d\theta} \Psi_q(\theta) = \frac{-\theta^{q-1} e^{-\theta}}{\Gamma(q)} < 0, \quad \forall q \in \mathbb{N}, \text{ and } \theta > 0. \tag{2.16}$$

The limiting behavior of $\Psi_q(\theta)$ is important. For a fixed q, the function $\Psi_q(\theta) \to 1$ as $\theta \to 0$ and $\Psi_q(\theta) \to 0$ as $\theta \to \infty$. While Ψ_q^{-1} still exists in these situations because Ψ_q is monotonically decreasing, for a given z the value $\Psi_q^{-1}(z)$ could be numerically very difficult to evaluate. To characterize the sets of θ and q that Ψ_q is (numerically) invertible, we define the *θ-admissible set* and the *q-admissible set*.

Definition 2.3.1 *The* θ-admissible set *and* q-admissible set *of the incomplete Gamma function are*

$$\begin{aligned} \Theta_q &\stackrel{def}{=} \{\theta \mid \varepsilon \le \Psi_q(\theta) \le 1 - \varepsilon\}, \\ \mathcal{Q}_\theta &\stackrel{def}{=} \{q \mid \varepsilon \le \Psi_q(\theta) \le 1 - \varepsilon\}, \end{aligned} \tag{2.17}$$

respectively, where $0 < \varepsilon < \frac{1}{2}$ *is a constant.*

More discussions of the incomplete Gamma function can be found in the supplementary material.

Remark 1 *In this thesis, we assume that QIS is noise-free, i.e., the only source of randomness is the truncated Poisson random variable. In real sensors, there will be readout noise, photo-response non-uniformity caused by conversion gain variation, dark count rate (a.k.a. dark current), optical crosstalk and electronic crosstalk. See [26] for details.*

Remark 2 *In Chapter 5, we slightly change the notation to avoid ambiguities with the color filter array model.*

3. QIS IMAGE RECONSTRUCTION

In this chapter, we tackle the QIS image reconstruction problem. First, we present an iterative approach for obtaining the ML solution in Section 3.1. We also derive a closed-form expression for the ML solution under certain conditions. Second, we present in Section 3.2 an iterative image reconstruction algorithm based on the MAP criterion. Compared to [22], this algorithm is more flexible where it can handle spatially-varying thresholds. Third, we present in Section 3.3 a fast and accurate reconstruction approach, which is based on the ML solution and a denoising step performed an appropriate transform domain.

3.1 Maximum Likelihood Estimation

Given $\mathcal{B} = \{B_{m,t} \mid m = 0, \ldots, M-1, \text{and } t = 0, \ldots, T-1\}$, MLE solves the following optimization problem:

$$\begin{aligned}
\widehat{\boldsymbol{c}} &\overset{(a)}{=} \underset{\boldsymbol{\theta}=\alpha\boldsymbol{G}\boldsymbol{c}}{\operatorname{argmax}} \prod_{t=0}^{T-1}\prod_{m=0}^{M-1} \mathbb{P}[B_{m,t}=1\,;\,\theta_m]^{b_{m,t}} \times \mathbb{P}[B_{m,t}=0\,;\,\theta_m]^{1-b_{m,t}} \\
&\overset{(b)}{=} \underset{\boldsymbol{\theta}=\alpha\boldsymbol{G}\boldsymbol{c}}{\operatorname{argmax}} \sum_{t=0}^{T-1}\sum_{m=0}^{M-1} \Big\{ b_{m,t}\log(1-\Psi_q(\theta_m)) + (1-b_{m,t})\log\Psi_q(\theta_m) \Big\} \\
&\overset{(c)}{=} \underset{\boldsymbol{\theta}=\alpha\boldsymbol{G}\boldsymbol{c}}{\operatorname{argmin}} \; F(\boldsymbol{\theta};\boldsymbol{B})
\end{aligned} \tag{3.1}$$

Here, the right hand side of (a) is the likelihood function of a Bernoulli random variable, (b) follows from taking the logarithm, and (c) follows from defining the negative log-likelihood function $F : \mathbb{R}^+ \times \{0,1\}^{MT} \to \mathbb{R}^+$ which is written as

$$F(\boldsymbol{\theta};\boldsymbol{B}) \overset{\text{def}}{=} -\sum_{t=0}^{T-1}\sum_{m=0}^{M-1} \Big\{ b_{m,t}\log(1-\Psi_q(\theta_m)) + (1-b_{m,t})\log\Psi_q(\theta_m) \Big\} \tag{3.2}$$

In [21], the authors prove that the log likelihood function is concave. Hence, $F(\boldsymbol{\theta},\boldsymbol{B})$ is convex in $\boldsymbol{\theta}$, and (3.1) is a convex optimization problem. However, for

general matrix $\boldsymbol{G}$, or equivalently a general interpolation kernel $\phi(x)$, the optimization problem is not separable in the variables $\{c_0, \ldots, c_{N-1}\}$. Hence, an iterative algorithm is required to solve it. On the other hand, for $\boldsymbol{G}$ defined in (2.11), the problem is separable in $\{c_0, \ldots, c_{N-1}\}$, and we can obtain a closed-form expression for the ML solution. In the next two subsections, we will discuss these two cases in more details.

3.1.1 ADMM Algorithm for Solving MLE

In this subsection we discuss how to solve the MLE problem in (3.1) using the alternating direction method of multipliers (ADMM) algorithm [101]. Our focus here is the modification required to accommodate the case of $q > 1$ and $\alpha > 1$ for the original ADMM algorithm presented in [22].

Inspecting (3.1), we note that it is an equality constrained optimization. Therefore, we can formulate its augmented Lagrangian function as

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{\theta}, \tilde{\boldsymbol{z}}) = F(\boldsymbol{\theta}; \boldsymbol{B}) - \tilde{\boldsymbol{z}}^T(\boldsymbol{\theta} - \alpha \boldsymbol{G}\boldsymbol{c}) + \frac{\rho}{2}\|\boldsymbol{\theta} - \alpha \boldsymbol{G}\boldsymbol{c}\|^2, \tag{3.3}$$

where $\tilde{\boldsymbol{z}} \in \mathbb{R}^m$ is the Lagrangian multiplier associated with the constraint $\boldsymbol{\theta} = \alpha \boldsymbol{G}\boldsymbol{c}$, and $\rho > 0$ is a non-negative scalar that control the strength of the quadratic penalty term. By completing squares and using the scaled Lagrangian multiplies $\boldsymbol{z} = \tilde{\boldsymbol{z}}/\rho$, the augmented Lagrangian can be rewritten as

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{z}) = F(\boldsymbol{\theta}; \boldsymbol{B}) + \frac{\rho}{2}\|\boldsymbol{\theta} - \alpha \boldsymbol{G}\boldsymbol{c} - \boldsymbol{z}\|^2 + \frac{\rho}{2}||\boldsymbol{z}||^2. \tag{3.4}$$

We can solve the optimization problem via an iterative approach

$$\boldsymbol{c}^{(k+1)} = \underset{\boldsymbol{c}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{c}, \boldsymbol{\theta}^{(k)}, \boldsymbol{z}^{(k)}), \tag{3.5a}$$

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{c}^{(k+1)}, \boldsymbol{\theta}, \boldsymbol{z}^{(k)}), \tag{3.5b}$$

$$\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} - \left(\boldsymbol{\theta}^{(k+1)} - \alpha \boldsymbol{G}\boldsymbol{c}^{(k+1)}\right). \tag{3.5c}$$

Since $F(\boldsymbol{\theta}; \boldsymbol{B})$ is convex in $\boldsymbol{\theta}$, convergence of (3.5a)-(3.5c) is guaranteed under appropriate conditions [101]. For notational simplicity, we will drop the iteration index on solving the $\boldsymbol{c}$-subproblem in (3.5a) and the $\boldsymbol{\theta}$-subproblem in (3.5b).

• ***c*-subproblem**: By defining the variable $\boldsymbol{c}_0 = \boldsymbol{\theta} - \boldsymbol{z}$ and dropping terms independent of $\boldsymbol{c}$, we can write the $\boldsymbol{c}$-subproblem as follows

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmin}} \quad \frac{\rho}{2}||\boldsymbol{c}_0 - \alpha \boldsymbol{G}\boldsymbol{c}||^2 \tag{3.6}$$

which is a quadratic optimization problem. This problem can be solved by setting the first derivative

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}} = -\alpha\rho \boldsymbol{G}^T(\boldsymbol{c}_0 - \alpha \boldsymbol{G}\boldsymbol{c}) \tag{3.7}$$

to zero, and solving for $\boldsymbol{c}$ to get

$$\boldsymbol{c} = \frac{1}{\alpha}(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{c}_0 \tag{3.8}$$

• ***θ*-subproblem**: This problem is more challenging as it involves the nonlinear incomplete Gamma function $\Psi_q(\theta_m)$. Substituting for the negative log-likelihood function $F(\boldsymbol{\theta}; \boldsymbol{B})$ from (3.2) and eliminating terms independent of $\boldsymbol{\theta}$, solving (3.5b) is equivalent to solving

$$\min_{\boldsymbol{\theta}} \sum_{t=0}^{T-1} \sum_{m=0}^{M-1} \left[\frac{\rho}{2}(\theta_m - d_m)^2 - (1 - b_{m,t})\log\left(\Psi_q(\theta_m)\right) - b_{m,t}\log\left(1 - \Psi_q(\theta_m)\right) \right], \tag{3.9}$$

where $\boldsymbol{d} = [d_0, \ldots, d_{M-1}]^T$ with $\boldsymbol{d} = \alpha\boldsymbol{G}\boldsymbol{c} + \boldsymbol{z}$. By defining the variable

$$S_m = \sum_{t=0}^{T-1} b_{m,t},$$

we can rewrite (3.9) as follows.

$$\min_{\boldsymbol{\theta}} \sum_{m=0}^{M-1} \left[\frac{\rho}{2}(\theta_m - d_m)^2 - (T - S_m)\log\left(\Psi_q(\theta_m)\right) - S_m\log\left(1 - \Psi_q(\theta_m)\right) \right]. \tag{3.10}$$

To solve (3.10), we recognize that it is a sum of M separable functions. Therefore, (3.10) is minimized when each individual term in the sum is minimized. The first order optimality returns us the following result.

Proposition 3.1.1 *The optimal solution θ_m of* (3.10) *satisfies the equations*

$$\rho\theta_m + \frac{e^{-\theta_m}\theta_m^{q-1}}{\Gamma(q)} \frac{T(1 - \psi_q(\theta_m)) - S_m}{\Psi_q(\theta_m)(1 - \Psi_q(\theta_m))} = \rho d_m, \quad \forall S_m \in \{0, 1, \ldots, T\}. \tag{3.11}$$

Proof By using the first order derivative of $\Psi_q(\theta_m)$ in (2.16), we can differentiate the m-th term in (3.10) and set the result to zero to yield (3.11). ■

From Proposition 3.1.1, it remains to solve (3.11). However, since (3.11) is a transcendental equation, we must adopt a numerical approach to solve the equation. Our proposed solution relies on building a look up table (offline) for $D+1$ values of d_m distributed uniformly in the interval $[d_{\min}, d_{\max}]$ with a step $\Delta d = (d_{\max} - d_{\min})/D$. Then, the solution at any value of d is obtained by a simple linear interpolation.

Remark: Because of the nonlinearity of the incomplete Gamma function $\Psi_q(\theta)$, when building the look up table a solution may lie in a region close to discontinuity. To mitigate this issue, we use a bisection to determine an approximate interval in which the solution must be contained.

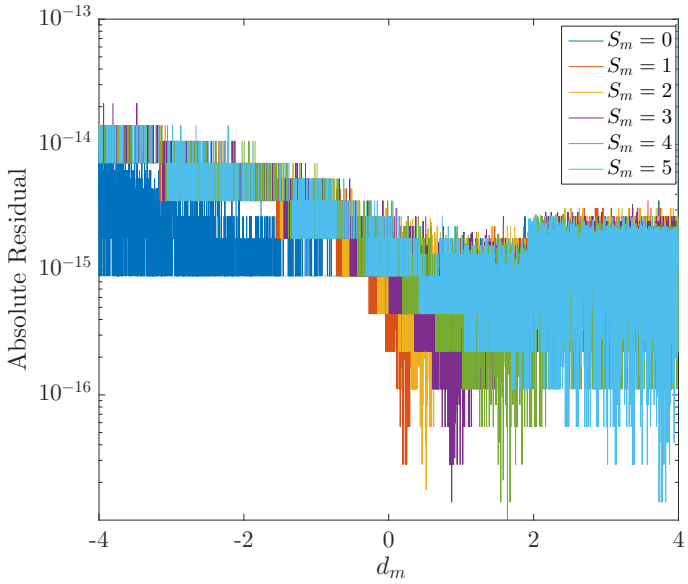


Fig. 3.1. Absolute Residual vs $d_m \in [-4, 4]$ after substituting with the obtained root in (3.11). The number of points is $D = 10^4$, and $T = 5$.

3.1.2 Closed-Form ML Expression for Box-car kernel

Under the box-car interpolation kernel assumption, the MLE problem can be simplified to obtain a closed-form expression for the ML solution. With the $\boldsymbol{G}$ defined in (2.11), we can partition $\mathcal{B}$ into N independent blocks $\{\mathcal{B}_1, \ldots, \mathcal{B}_N\}$ where each block is

$$\mathcal{B}_n \stackrel{\text{def}}{=} \{B_{Kn+k,t} \mid k = 0, \ldots, K-1, t = 0, \ldots, T-1\}. \tag{3.12}$$

In addition, the constraint $\boldsymbol{\theta} = \alpha \boldsymbol{G}\boldsymbol{c}$ can be rewritten as

$$\theta_{Kn+k} = \frac{\alpha c_n}{K}, \quad \forall n \in \{0, \ldots, N-1\} \tag{3.13}$$

Then, the pixel $\widehat{c}_n$ can be estimated according to the following proposition.

Proposition 3.1.2 (Closed-form ML Estimate) *For $\phi(x)$ defined as box-car kernel, the solution of the MLE in* (3.1) *is given by*

$$\widehat{c}_n = \frac{K}{\alpha} \Psi_q^{-1}\left(1 - \frac{S_n}{KT}\right), \quad \forall n \in \{0, \ldots, N-1\} \tag{3.14}$$

where $S_n \stackrel{def}{=} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} B_{Kn+k,t}$ is the sum of bits in the n-th block $\mathcal{B}_n$.

Proof See Appendix C.1. ■

It would be instructive to illustrate Proposition 3.1.2 using a figure. Figure 3.2 shows the case when $T = 1$, i.e., a single exposure, and $K = 16$. The 1-bit measurements are first averaged to compute the number of ones within a block of size K. Then, applying the inverse incomplete Gamma function $\Psi_q^{-1}(\cdot)$ and a scaling constant K/α we obtain the solution $\widehat{c}_n$.

3.2 Maximum-A-Posterior Solution

Since the ML solution depends solely on input random data, it contains some randomness which makes the ML solution noisy. This noise is exacerbated when the combined spatial-temporal oversampling L is not large enough because the noise

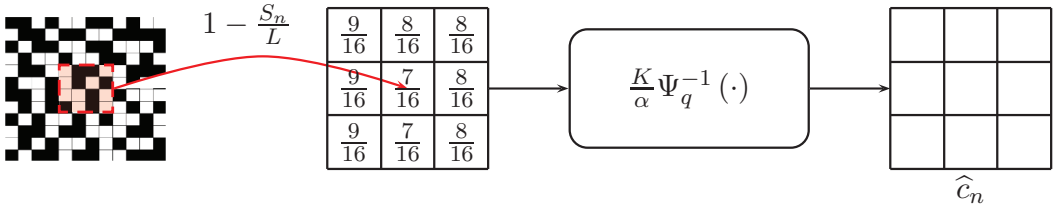


Fig. 3.2. Pictorial interpretation of Proposition 3.1.2: Given an array of 1-bit measurements (black = 0, white = 1), we compute the number of ones within a block of size K. Then the solution of the MLE problem in (C.3) is found by applying an inverse incomplete Gamma function $\Psi_q^{-1}(\cdot)$ and a scaling factor K/α.

variance will be significant for small L. This problem can be alleviated by using our prior knowledge of the attributes of the output image within the Maximum-A-Posterior (MAP) framework [61]. Denote the negative logarithm of the prior function as $g(\boldsymbol{c}) = -\log(p(\boldsymbol{c}))$, the MAP estimation problem can be written as

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{\theta}=\alpha\boldsymbol{G}\boldsymbol{c}}{\operatorname{argmin}} \quad F(\boldsymbol{\theta};\boldsymbol{B}) + g(\boldsymbol{c}) \tag{3.15}$$

As for choosing the prior $g(\boldsymbol{c})$, there are many options that include Gaussian and non Gaussian Markov random fields [61], sparsity-based priors [102], data-driven priors learned by neural networks [103], denoising-based priors in the plug-and-play framework [104]. Here, we use the anisotropic total variation prior [22] for simplicity, where extension to other priors is straightforward. Denote by $\boldsymbol{D}$ the first order finite difference operator. Hence, we cab formulate the MAP estimation problem with total variation prior as follows.

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{\theta}=\alpha\boldsymbol{G}\boldsymbol{c}}{\operatorname{argmin}} \; F\left(\boldsymbol{\theta};\boldsymbol{B}\right) + \lambda||\boldsymbol{D}\boldsymbol{c}||_1, \tag{3.16}$$

which can be rewritten as follows

$$\widehat{\boldsymbol{c}} = \underset{\substack{\boldsymbol{\theta}=\alpha\boldsymbol{G}\boldsymbol{c} \\ \boldsymbol{v}=\boldsymbol{D}\boldsymbol{c}}}{\operatorname{argmin}} \; F\left(\boldsymbol{\theta};\boldsymbol{B}\right) + \lambda||\boldsymbol{v}||_1. \tag{3.17}$$

By completing squares as we did before, we can write the augmented Lagrangian as follows.

$$\mathcal{L}(\boldsymbol{c},\boldsymbol{\theta},\boldsymbol{v},\boldsymbol{z},\boldsymbol{r}) = F(\boldsymbol{\theta};\boldsymbol{B}) + \lambda||\boldsymbol{v}||_1 + \frac{\rho}{2}||\boldsymbol{\theta} - \alpha\boldsymbol{G}\boldsymbol{c} - \boldsymbol{z}||^2 - \frac{\rho}{2}||\boldsymbol{z}||^2 + \frac{\gamma}{2}||\boldsymbol{v} - \boldsymbol{D}\boldsymbol{c} - \boldsymbol{r}||^2 - \frac{\gamma}{2}||\boldsymbol{r}||^2 \tag{3.18}$$

where $\boldsymbol{z} \in \mathbb{R}^M$ and $\boldsymbol{r} \in \mathbb{R}^N$ are the scaled Lagrangian variables associated with the constraints $\boldsymbol{\theta} = \alpha\boldsymbol{G}\boldsymbol{c}$ and $\boldsymbol{v} = \boldsymbol{D}\boldsymbol{c}$, respectively. ρ and γ are non-negative weights the control the power of the quadratic penalty terms. Using the ADMM framework, we can minimize the augmented Lagrangian by solving the following sequence of subproblems.

$$\boldsymbol{c}^{(k+1)} = \underset{\boldsymbol{c}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{c},\boldsymbol{\theta}^{(k)},\boldsymbol{v}^{(k)},\boldsymbol{z}^{(k)},\boldsymbol{r}^{(k)}), \tag{3.19a}$$

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{c}^{(k+1)},\boldsymbol{\theta},\boldsymbol{v}^{(k)},\boldsymbol{z}^{(k)},\boldsymbol{r}^{(k)}), \tag{3.19b}$$

$$\boldsymbol{v}^{(k+1)} = \underset{\boldsymbol{v}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{c}^{(k+1)},\boldsymbol{\theta}^{(k+1)},\boldsymbol{v},\boldsymbol{z}^{(k)},\boldsymbol{r}^{(k)}), \tag{3.19c}$$

$$\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} - \left(\boldsymbol{\theta}^{(k+1)} - \alpha\boldsymbol{G}\boldsymbol{c}^{(k+1)}\right). \tag{3.19d}$$

$$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \left(\boldsymbol{v}^{(k+1)} - \boldsymbol{D}\boldsymbol{c}^{(k+1)}\right). \tag{3.19e}$$

From the convexity of $F(\boldsymbol{\theta};\boldsymbol{B})$ and the total variation prior, convergence of (3.19a)-(3.19e) is generally guaranteed [101]. For notational simplicity, we will drop the superscripts on presenting the solution of each subproblem.

- ***c*-subproblem**: By defining the variables $\boldsymbol{c}_0 = \boldsymbol{\theta} - \boldsymbol{z}$ and $\boldsymbol{c}_1 = \boldsymbol{v} - \boldsymbol{r}$, and dropping terms independent of $\boldsymbol{c}$, we can write the $\boldsymbol{c}$-subproblem as follows.

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmin}} \quad \frac{\rho}{2}||\boldsymbol{c}_0 - \alpha\boldsymbol{G}\boldsymbol{c}||^2 + \frac{\gamma}{2}||\boldsymbol{c}_1 - \boldsymbol{D}\boldsymbol{c}||^2 \tag{3.20}$$

which can be solved by setting the first derivative to zero and rearranging the terms to get

$$\widehat{\boldsymbol{c}} = \left(\rho\alpha^2\boldsymbol{G}^T\boldsymbol{G} + \gamma\boldsymbol{D}^T\boldsymbol{D}\right)^{-1}\left(\rho\alpha\boldsymbol{G}^T\boldsymbol{c}_0 + \gamma\boldsymbol{D}^T\boldsymbol{c}_1\right) \tag{3.21}$$

Since the matrix $\rho\alpha^2\boldsymbol{G}^T\boldsymbol{G} + \gamma\boldsymbol{D}^T\boldsymbol{D}$ is circulant as proved in [22], the inversion can be implemented in the Foruier domain to improve computational efficiency.

• $\boldsymbol{\theta}$**-subproblem**: By defining the variable $\boldsymbol{d} = \alpha \boldsymbol{G}\boldsymbol{c} + \boldsymbol{z}$, we can write the $\boldsymbol{\theta}$-subproblem as follows.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \ \frac{\rho}{2}||\boldsymbol{\theta} - \boldsymbol{d}||^2 + F(\boldsymbol{\theta}; \boldsymbol{B}) \tag{3.22}$$

which is the same as the $\boldsymbol{\theta}$-subproblem (3.9) in the ML solution. Hence, it can be solved in the same way.

• $\boldsymbol{v}$**-subproblem**: By defining the variable $\boldsymbol{v}_0 = \boldsymbol{D}\boldsymbol{c} + \boldsymbol{r}$, we can write the $\boldsymbol{v}$-subproblem as follows

$$\widehat{\boldsymbol{v}} = \underset{\boldsymbol{v}}{\operatorname{argmin}} \ \ \lambda||\boldsymbol{v}||_1 + \frac{\gamma}{2}||\boldsymbol{v} - \boldsymbol{v}_0||^2 \tag{3.23}$$

Applying the shrinkage formula [61], the solution is

$$\widehat{\boldsymbol{v}} = \operatorname{sign}(v_0) \max(|v_0| - \lambda/\gamma, 0) \tag{3.24}$$

3.2.1 The Plug-and-Play Algorithm [104]

The choice of prior function $g(\boldsymbol{c})$ affects the reconstruction quality significantly. To choose priors stronger than the total-variation prior, we use the Plug-and-Play algorithm [104]. In this approach, the authors noticed that for a general prior $g(\boldsymbol{c})$, the $\boldsymbol{v}$-subproblem is equivalent to a denoising problem of a signal impaired with additive white Gaussian noise having standard deviation of $sqrt\lambda/\gamma$. Hence, the $\boldsymbol{v}$-subproblem can be solved by any standard Gaussian denoiser such as BM3D [105], Monte-Carlo Non-Local Means [106], a neural network denoiser [107, 108], or even a combination of Gaussian denoisers [109].

In our problem settings, using the Plug-and-Play approach will only change the $\boldsymbol{v}$-subproblem to be just a denoising step as follows

$$\widehat{\boldsymbol{v}} = \mathcal{D}_\sigma(\boldsymbol{v}_0), \tag{3.25}$$

where $\mathcal{D}$ is the selected denoiser, and $\sigma = \sqrt{\lambda/\gamma}$ is the noise level. All other subproblems will be the same. We can also use a sequence of decreasing noise levels to guarantee convergence as proved in [110].

3.3 Transform-Denoise Pipeline

Looking back at the two previous sections, we proposed 1) an iterative image reconstruction approach for ML or MAP estimation framework, and 2) a fast image reconstruction approach based on ML closed-form expression under box-car interpolation kernel assumption. Each approach has its pros and cons. The iterative approach can get a clean reconstructed image if a suitable prior is used in the MAP framework; however, it is impractical for ultra-fast applications. On the other hand, the closed-form expression gives an ML estimate in very short time with low computational complexity; however, the reconstructed image is noisy especially for small combined oversampling factor L. In this section, we propose an approach that combines the pros of both techniques.

3.3.1 Related Work in the Literature

The proposed algorithm belongs to a family of methods we call the *transform-denoise* methods. The idea of transform-denoise is similar to what we do here: Transform the random variable using a variance stabilizing transform, then denoise using an off-the-shelf image denoiser. Among the existing transform-denoise methods, perhaps the most notable work is the one by Makitalo and Foi [111], where they considered the optimal inverse of the Anscombe Transform for the case of Poisson-Gaussian random variables. A more recent work by the same research group [112] showed that it is possible to boost the denoising performance by applying the transform-denoise iteratively. We should also mention the work by Foi [113], which considered the modeling and transformation for clipped noisy images. The problem setting of that work is for conventional sensors. However, the underlying principle using the transform-denoise approach is similar to that of QIS.

The closed-form ML expression in subsection 3.1.2 is based on the box-car iterpolation kernel assumption (i.e., $\boldsymbol{G}$ defined in (2.11)). Under this assumption, summing of the Bernoulli random variables can be thought of performing a spatial-temporal

"binning" of the pixels. Binning is a common technique in restoring images from Poisson noise, especially when the signal-to-noise ratio is low [114–116]. Binning can also be applied together with transform-denoise, e.g., in [112], to achieve improved results. For QIS, the result of binning is different from that of the Poisson noise, for the sum of QIS bits leads to a binomial random variables whereas the sum of Poisson noise leads to a Poisson random variable.

3.3.2 Binomial Anscombe Transform

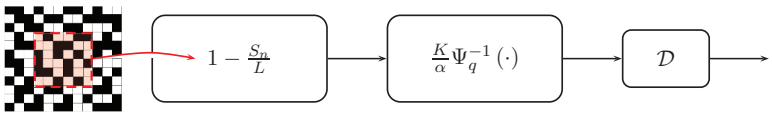
The MLE solution $\widehat{\boldsymbol{c}} = [\widehat{c}_0, \ldots, \widehat{c}_{N-1}]^T$ computed through Proposition 3.1.2 is noisy. The reason is that for a relatively small K and T, the randomness in the 1-bit measurement has not yet been eliminated by the summation in S_n. Therefore, in order to improve the image quality, additional steps must be taken to improve the smoothness of the image.

At the first glance, this question seems easy because if one wants to mitigate the noise in $\widehat{\boldsymbol{c}}$, then directly applying an image denoising algorithm $\mathcal{D}$ to $\widehat{\boldsymbol{c}}$ would be sufficient, e.g., Figure 3.3(a). However, a short afterthought will suggest that such approach is invalid for the following reason. For the majority of image denoising algorithms in the literature, the noise is assumed to be independently and identically distributed (i.i.d.) Gaussian. In other words, the variance of the noise should be spatially invariant. However, the resulting random variable $\widehat{\boldsymbol{c}}$ does not have this property.

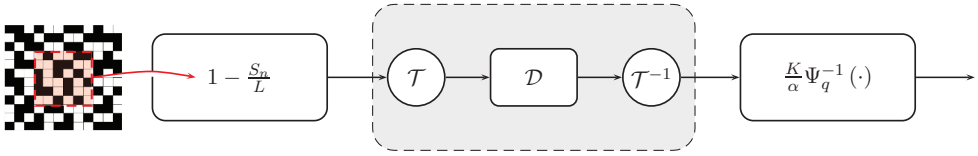
Our proposed solution is to apply an image denoiser *before* the inverse incomplete Gamma function as shown in Figure 3.3(b). Besides the order of denoising and Gamma function, we also add a pair of nonlinear transforms $\mathcal{T}$ and $\mathcal{T}^{-1}$ before and after the denoiser $\mathcal{D}$. The reasons of these two changes are based on the following observations.

Observation 3.3.1 *Under box-car kernel assumption, the random variables*

$$\{B_{Kn+k,t} \,|\, k = 0, \ldots, K-1, \text{ and } t = 0, \ldots, T-1\}$$



(a) Conventional method



(b) Proposed method

Fig. 3.3. Two possible ways of improving image smoothness for QIS. (a) The **conventional** approach denoises the image *after* $\widehat{c}_n$ is computed. (b) The **proposed** approach: Apply the denoiser *before* the inverse incomplete Gamma function, together with a pair of Anscombe transforms $\mathcal{T}$. The symbol $\mathcal{D}$ in this figure denotes a generic Gaussian noise image denoiser.

are i.i.d. Bernoulli of equal probability $\mathbb{P}[B_{Kn+k,t} = 1] = 1 - \Psi_q\left(\frac{\alpha c_n}{K}\right)$ *for* $k = 0, \ldots, K-1$ *and* $t = 0, \ldots, T-1$.

The proof of Observation 3.3.1 follows immediately from the matrix $\boldsymbol{G}$ defined in (2.11) which corresponds to the box-car kernel. We can divide the M jots into N groups each having $K \times T$ entries. Within the group, the 1-bit measurements are all generated from the same pixel c_n.

The consequence of Observation 3.3.1 is that for a sequence of i.i.d. Bernoulli random variables, the sum is a Binomial random variable. This is described in Observation 3.3.2.

Observation 3.3.2 *If* $\{B_{Kn+k,t}\}$ *are i.i.d. Bernoulli random variables with probability* $\mathbb{P}[B_{Kn+k,t} = 1] = 1 - \Psi_q\left(\frac{\alpha c_n}{K}\right)$ *for* $k = 0, \ldots, K-1$ *and* $t = 0, \ldots, T-1$, *then the sum* S_n *defined in* (C.2) *is a Binomial random variable with mean and variance*

$$\mathbb{E}[S_n] = L\left(1 - \Psi_q\left(\frac{\alpha c_n}{K}\right)\right), \qquad \mathrm{Var}[S_n] = L\Psi_q\left(\frac{\alpha c_n}{K}\right)\left(1 - \Psi_q\left(\frac{\alpha c_n}{K}\right)\right).$$

Observation 3.3.2 is a classic result in probability. The mean of the Bernoulli random variables is specified by the incomplete Gamma function $\Psi_q\left(\frac{\alpha c_n}{K}\right)$, which approaches 1 as K increases. Thus, for fixed T, the probability $1-\Psi_q\left(\frac{\alpha c_n}{K}\right) \to 0$ as $K \to \infty$. When this happens, the binomial random variable S_n can be approximated by a Poisson random variable with mean $L\left(1-\Psi_q\left(\frac{\alpha c_n}{K}\right)\right)$ [117]. However, as T also grows, the binomial random variable S_n can be further approximated by a Gaussian random variable due to the Central Limit Theorem. Therefore, for a reasonably large K and T, the resulting random variable S_n is approximately Gaussian.

The variance of this approximated Gaussian is, however, not constant. The variance changes across different locations n because $\mathrm{Var}[S_n]$ is a function of c_n. Therefore, if we want to apply a conventional image denoiser (which assumes i.i.d. Gaussian noise) to smooth S_n, we must first make sure that the noise variance is spatially invariant. The technique used to accomplish this goal is called the variance stabilizing transform [118]. In this paper, we use a specific variance stabilizing transform known as the Anscombe Transform [119]. Anscombe Transform is best known in the image processing literature for Poisson denoising, where one transforms an observed Poisson data to approximately Gaussian with equal variance [111]. For binomial random variables S_n, the Anscombe Transform and its property are given in Theorem 3.3.1.

Theorem 3.3.1 (Anscombe Transform for Binomial Random Variables) *Let S_n be a binomial random variable with parameters (L, p_n), where $p_n = 1-\Psi_q\left(\frac{\alpha c_n}{K}\right)$ and $L = KT$. Define the Anscombe Transform of S_n as a function $\mathcal{T}: \{0,\ldots,L\} \to \mathbb{R}$ such that*

$$Z_n = \mathcal{T}(S_n) \overset{def}{=} \sqrt{L+\frac{1}{2}}\sin^{-1}\left(\sqrt{\frac{S_n+\frac{3}{8}}{L+\frac{3}{4}}}\right). \tag{3.26}$$

Then, the variance of Z_n is $\mathrm{Var}[Z_n] = \frac{1}{4} + \mathcal{O}(L^{-2})$ for all n.

Proof The proof of Theorem 3.3.1 is given in the Appendix. It is a simplified version of a technical report by Brown et al. [120]. The original paper by Anscombe [119] also contains a sketch of the proof. However, the sketch is rather brief and we believe that a complete derivation would make this thesis self-contained. ■

The implication of Theorem 3.3.1 is that regardless of the location n, the transformed random variable Z_n has a constant variance $\frac{1}{4}$ when L is large. Therefore, the noise variance is now location independent and hence a standard i.i.d. Gaussian denoiser can be used.

Example 2 *To provide readers a demonstration of the effectiveness of Theorem 3.3.1, we consider a checkerboard image of* $N = 64$ *pixels with intensity levels* $c_0, \ldots, c_{N-1}$. *The* n*-th pixel* c_n *generates* $K = 100$ *binary quantized Poisson measurements*

$$\{B_{Kn}, \ldots, B_{Kn+(K-1)}\}$$

using $\alpha = 100$, $q = 1$, $T = 1$ *(So* $L = 100$*). From each of these* K *measurements we sum to obtain a binomial random variable* $S_n = \sum_{k=0}^{K-1} B_{Kn+k}$. *We then compute the variance of* $\mathrm{Var}[S_n]$ *and* $\mathrm{Var}[\mathcal{T}(S_n)]$ *using* 10^4 *independent Monte Carlo trials. The results are shown in Figure 3.4, where we observe that* $\mathrm{Var}[S_n]$ *varies with the location* n, *and* $\mathrm{Var}[\mathcal{T}(S_n)]$ *is nearly constant for all* n.

Remark. The inverse Anscombe Transform is

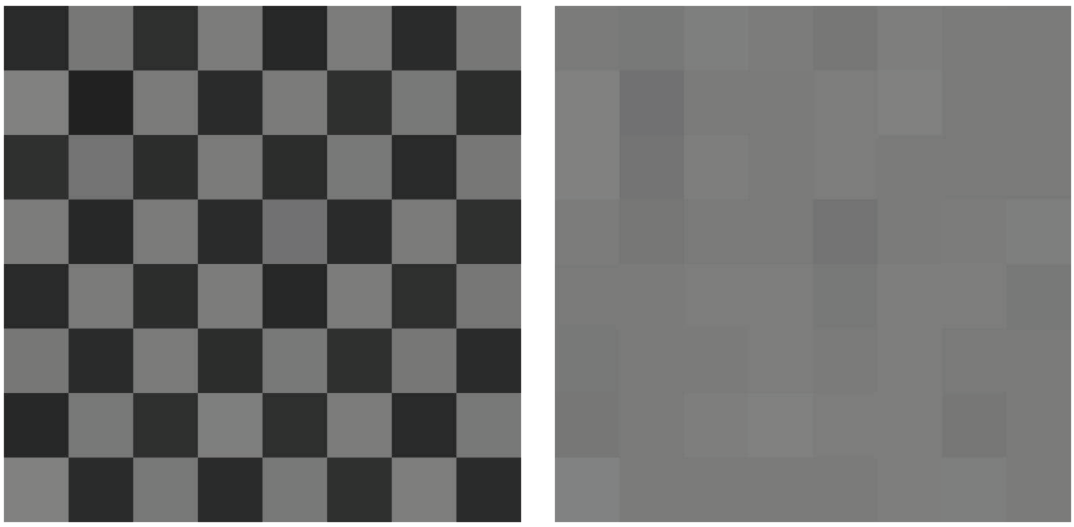
$$S_n = \mathcal{T}^{-1}(Z_n) = \left(L + \frac{3}{4}\right) \sin^2\left(\frac{Z_n}{\sqrt{L + \frac{1}{2}}}\right) - \frac{3}{8}, \tag{3.27}$$

which we call it the *algebraic inverse*. Another possible inverse of the Anscombe Transform is the *asymptotic unbiased inverse* [119], defined as

$$S_n = \mathcal{T}^{-1}_{\text{unbias}}(Z_n) = \left(1 + \frac{1}{2L}\right)^{-1} \left[\left(L + \frac{3}{4}\right) \sin^2\left(\frac{Z_n}{\sqrt{L + \frac{1}{2}}}\right) - \frac{1}{8}\right]. \tag{3.28}$$

For large L, the difference between the asymptotic unbiased inverse and the algebraic inverse is small.

Example 3 *Table 3.1 shows the PSNR values of the reconstructed images using the algebraic inverse and the asymptotic unbiased inverse. In this experiment, we consider 10 standard images commonly used in the image processing literature: Baboon,*



$\mathrm{Var}[S_n]$ $\mathrm{Var}[\mathcal{T}(S_n)]$

Fig. 3.4. Illustration of Anscombe Transform. Both sub-figures contain $N = 64$ (8×8) pixels $c_0, \ldots, c_{N-1}$. For each pixel we generate 100 binary Poisson measurements and sum to obtain binomial random variables $S_0, \ldots, S_{N-1}$. We then calculate the variance of each S_n. Note the constant variance after the Anscombe Transform.

Barbara, Boat, Bridge, Couple, Hill, House, Lena, Man and Peppers. The sizes of the images are either 256×256 *or* 512×512. *For each image, we set* $T = 1$, $q = 1$, *and* $\alpha = K$, *and vary* $K = \{1, 4, 9, 16, 25, 36, 49, 64\}$. *The results in Table 3.1 indicate that* $\mathcal{T}^{-1}_{\text{unbias}}$ *is consistently better than* $\mathcal{T}^{-1}$ *for* $K > 1$, *although the difference diminishes as* K *grows.*

Table 3.1.
PSNR values using algebraic inverse $\mathcal{T}^{-1}$ and asymptotic unbiased inverse $\mathcal{T}^{-1}_{\text{unbias}}$. The results are averaged over 10 standard images. In this experiment, we set $T = 1$, $q = 1$, and $\alpha = K$.

K	1	4	9	16	25	36	49	64
$\mathcal{T}^{-1}$	20.51	23.08	25.00	26.47	27.49	28.40	29.09	29.71
$\mathcal{T}^{-1}_{\text{unbias}}$	19.43	23.64	25.30	26.62	27.57	28.45	29.12	29.73

4. OPTIMAL THRESHOLD DESIGN: THEORY AND PRACTICE

In this chapter, we study the QIS threshold design problem. In our theoretical derivations, we focus on the ML estimate as it provides closed-form expressions under box-car interpolation kernel assumption. We start in Section 4.1 by studying a theoretical oracle scenario when the ground truth is assumed to be known. This study form the basis of our subsequent discussions of the threshold update scheme in Section 4.2 where we tackle the practical case of unknown ground truth.

4.1 Optimal Threshold: Theory

In this section, we tackle the oracle scenario where the ground truth is given. We start by obtaining in subsection 4.1.1 a closed-form expression of the the ML estimate's SNR in terms of the incomplete Gamma function. Then, we derive in subsection 4.1.2 an expression for the optimal "oracle" threshold that maximizes the SNR given the ground truth. This oracle threshold provides us with intuition how to tackle the realistic case when the ground truth is unknown.

4.1.1 Signal-to-Noise Ratio of ML Estimate

In order to determine the optimal threshold, we need to quantify the performance of the ML estimate. The performance metric we use is the signal-to-noise ratio of the ML estimate at every pixel $\widehat{c}_n$. Considering each $\widehat{c}_n$ individually is allowed here because they are independently determined according to (3.14). For notation simplicity we drop the subscript n in the subsequent discussions.

Definition 4.1.1 *The signal-to-noise ratio (SNR) of the ML estimate $\widehat{c}$ is defined as*

$$\mathrm{SNR}_q(c) \stackrel{def}{=} 10\log_{10}\frac{c^2}{\mathbb{E}[(\widehat{c}-c)^2]}, \tag{4.1}$$

where the expectation is taken over the probability mass function of the binary measurements in (2.15).

The difficulty of working with $\mathrm{SNR}_q(c)$ is that it does not have a simple closed-form expression. In view of this, Lu [64] showed that the SNR is asymptotically linear to the log of the Fisher Information.

Proposition 4.1.1 *As $KT \to \infty$,*

$$\mathrm{SNR}_q(c) \approx 10\log_{10}\left(c^2 I_q(c)\right) + 10\log_{10} KT, \tag{4.2}$$

where $I_q(c) \stackrel{def}{=} \mathbb{E}_B\left[\frac{-\partial^2}{\partial c^2}\log \mathbb{P}(B=b;\theta)\right]$ is the Fisher Information measuring the amount of information that the random variable B carries about the unknown value c.

Proof See [64]. ■

While the asymptotic result shown in Proposition 4.1.1 has significantly simplified the SNR, we still need to determine the Fisher Information. The following proposition gives a new result of the Fisher Information with arbitrary q.

Proposition 4.1.2 *The Fisher Information $I_q(c)$ of the probability mass function in* (2.15) *under a threshold q is:*

$$I_q(c) = \left(\frac{\alpha}{K}\right)^2 \frac{e^{-2\left(\frac{\alpha c}{K}\right)}\left(\frac{\alpha c}{K}\right)^{2q-2}}{\Gamma^2(q)\Psi_q\left(\frac{\alpha c}{K}\right)\left(1-\Psi_q\left(\frac{\alpha c}{K}\right)\right)}. \tag{4.3}$$

Proof See Appendix C.3. ■

Substituting (4.3) into (4.2), we observe that the SNR can be approximated as

$$\mathrm{SNR}_q(c) \approx 10\log_{10}\frac{KTe^{-2\left(\frac{\alpha c}{K}\right)}\left(\frac{\alpha c}{K}\right)^{2q}}{\Gamma(q)^2\Psi_q\left(\frac{\alpha c}{K}\right)\left(1-\Psi_q\left(\frac{\alpha c}{K}\right)\right)}, \tag{4.4}$$

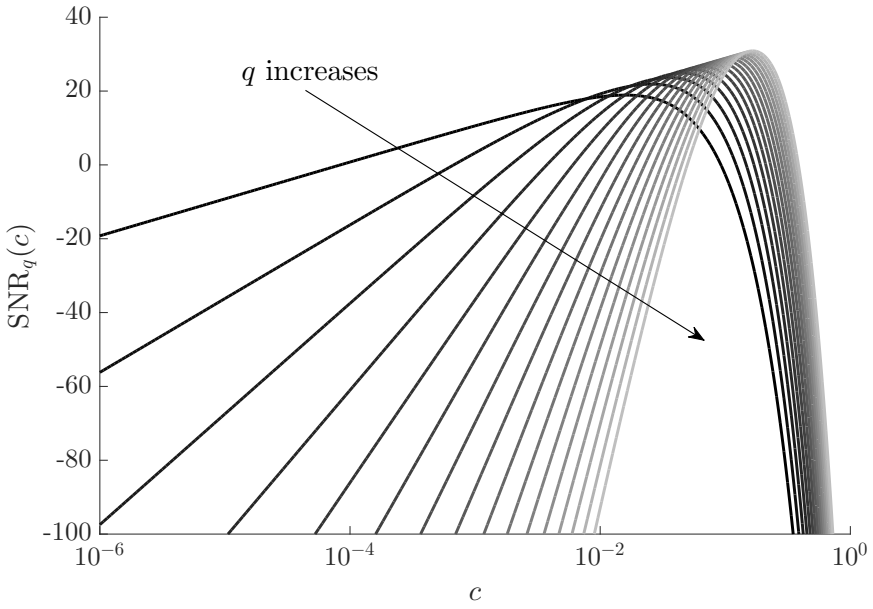


Fig. 4.1. $\mathrm{SNR}_q(c)$ for different thresholds $q \in \{1, \ldots, 16\}$. In this experiment, we set $\alpha = 400$, $K = 4$, and $T = 30$. For fixed q, $\mathrm{SNR}_q(c)$ is always a convex function.

which is characterized by the unknown pixel value c, the threshold q, the spatial oversampling ratio K and the number of temporal measurements T. To understand the behavior of (4.4), we show in Figure 4.1 $\mathrm{SNR}_q(c)$ as a function of c for different thresholds $q \in \{1, \ldots, 16\}$. For a fixed q, $\mathrm{SNR}_q(c)$ is a convex function with a unique maximum. The goal of optimal threshold design is to determine a q which maximizes $\mathrm{SNR}_q(c)$ for a fixed c.

Remark 3 *The* $\mathrm{SNR}_q(c)$ *in* (4.4) *can also be derived from a concept in the device literature called the* exposure-referred SNR *[26]. See Supplementary Material for discussions.*

4.1.2 Oracle Threshold

We now discuss the optimal threshold design in the oracle setting. We call the result oracle because the optimal threshold depends on the unknown pixel intensity c. The practical threshold design scheme will be discussed in Section 4.2.

Using the definition of the signal-to-noise ratio, the optimal threshold is determined by maximizing $\mathrm{SNR}_q(c)$ with respect to q:

$$q^* = \underset{q\in\mathbb{N}}{\operatorname{argmax}} \ \ \mathrm{SNR}_q(c) = \underset{q\in\mathbb{N}}{\operatorname{argmax}} \ \ \log(c^2 I_q(c)). \tag{4.5}$$

The second equality follows from Proposition 4.1.1. Substituting (4.3) yields an expression of the right hand side of (4.5). To further simplify the expression we derive the following lower bound.

Proposition 4.1.3 *The function* $\log(c^2 I_q(c))$ *is lower bounded as follows.*

$$\log(c^2 I_q(c)) \geq \underbrace{2\left(\log 2 - \frac{\alpha c}{K} + q\log\frac{\alpha c}{K} - \log\Gamma(q)\right)}_{\stackrel{def}{=} L_q(c)}.$$

Proof See Appendix C.4. ■

Using this lower bound, we can derive the optimal threshold q as follows [1].

Proposition 4.1.4 *The optimal threshold* $q^*(c)$ *is*

$$q^*(c) = \underset{q\in\mathbb{N}}{\mathit{argmax}} \ L_q(c) = \left\lfloor \frac{\alpha c}{K} \right\rfloor + 1, \tag{4.6}$$

where $\lfloor\cdot\rfloor$ *denotes the flooring operator that returns the largest integer smaller than or equal to the argument.*

Proof See Appendix C.5. ■

The result of Proposition 4.1.4 is important as it states that the oracle threshold is *exactly the same* as the light intensity $\alpha c/K$. The flooring operation and the addition of a constant 1 are not crucial here because they are only used to ensure that q is an integer. In [62], a special where $\alpha = 1$ was demonstrated experimentally. Proposition 4.1.4 now provides a theoretical justification.

[1] Straightly speaking, the result shown in Proposition 4.1.4 is a "near-optimal" result because we are minimizing the lower bound. From our experience, the gap between the near-optimality and the exact optimality is typically insignificant.

4.2 Optimal Threshold: Practice

The oracle threshold derived in the previous section provides a theoretical foundation but is practically infeasible as it requires knowledge of the ground truth c. In this section, we present an alternative solution by relaxing the optimality criteria. Our strategy is to consider a set of thresholds which are close to the oracle threshold $q^*(c)$, and show that they are asymptotically unbiased when the number of observed bits approaches infinity (subsection 4.2.1). This result will allow us to characterize the estimate $\widehat{c}$ (subsection 4.2.2). We will then show that there exists a phase transition region where the asymptotic unbiasedness is maintained as q stays within a certain range around $q^*(c)$, and is lost rapidly as q falls outside this range (subsections 4.2.3 and 4.2.4). Based on these observations, we will present a practical threshold update scheme (subsection 4.2.5). Finally, we discuss in subsection 4.2.6 how the threshold adaptation helps in extending the sensor's dynamic range for high dybnamic range imaging followed by some hardware considerations in subsection 4.2.7

4.2.1 Asymptotic Unbiasedness

In order to derive an alternative threshold that does not require the ground truth, we start by reconsidering the ML estimate $\widehat{c}$ in Proposition 3.1.2. For a spatial-temporal block $\mathcal{B} = \{B_{k,t} \,|\, 0 \leq k < K-1, 0 \leq t < T-1\}$, the ML estimate $\widehat{c}$ satisfies the condition

$$\Psi_q\left(\frac{\alpha\widehat{c}}{K}\right) = 1 - \frac{S}{KT}, \tag{4.7}$$

where $S = \sum_{k,t} B_{k,t}$ is the sum of bits in $\mathcal{B}$. The right hand side of this equation is an important quantity. We denote it as

$$\gamma_q(c) \stackrel{\text{def}}{=} 1 - \frac{S}{KT}. \tag{4.8}$$

In the device literature (e.g., [26]), the term $1 - \gamma_q(c)$ is known as the *bit-density* as it is the proportion of ones in $\mathcal{B}$. Note that $\gamma_q(c)$ is a random variable because S is

the sum of KT i.i.d. random binary bits. Therefore, if we want to understand (4.7), we must first derive the the mean and variance of $\gamma_q(c)$.

Proposition 4.2.1 *The mean and variance of* $\gamma_q(c)$ *are*

$$\begin{aligned}\mathbb{E}[\gamma_q(c)] &= \Psi_q\left(\frac{\alpha c}{K}\right), \quad \textit{and} \\ \mathrm{Var}[\gamma_q(c)] &= \frac{1}{KT}\Psi_q\left(\frac{\alpha c}{K}\right)\left[1-\Psi_q\left(\frac{\alpha c}{K}\right)\right], \end{aligned} \tag{4.9}$$

respectively.

Proof See Appendix C.6. ■

We can now look at the asymptotic behavior of $\gamma_q(c)$ to see if it offers any insight about the optimal threshold. Applying the strong law of large number to S/KT, we can show that as $KT \to \infty$,

$$\gamma_q(c) = 1 - S/KT \overset{a.s.}{\to} 1 - \mathbb{E}[B_{k,t}] = \Psi_q(\alpha c/K). \tag{4.10}$$

Going back to (4.7)-(4.8), the ML estimate $\widehat{c}$ should have the expectation:

$$\mathbb{E}[\widehat{c}] \overset{(a)}{=} \frac{K}{\alpha}\mathbb{E}\left[\Psi_q^{-1}\left(\gamma_q(c)\right)\right] \overset{(b)}{\to} \frac{K}{\alpha}\Psi_q^{-1}\Psi_q\left(\frac{\alpha c}{K}\right) \overset{(c)}{=} c. \tag{4.11}$$

where (a) follows from the definition of $\widehat{c}$, (b) follows from (4.10), and (c) holds because Ψ_q and Ψ_q^{-1} cancels each other.

What is the implication of (4.11)? It shows that the ML estimate $\widehat{c}$ is asymptotically unbiased. That is, as the number of independent measurements grows, the estimate $\widehat{c}$ approaches to the ground truth c. In other words, as long as KT is large enough, the random variable $\widehat{c}$ would be an accurate estimate of the ground truth. How can this be used to determine the threshold q? Let us look at $\mathcal{Q}_\theta$.

4.2.2 Set of Admissible Thresholds $\mathcal{Q}_\theta$

The result in (4.7)-(4.11) shows that for a given S (or equivalently $\gamma_q(c)$), the ML estimate can be found by

$$\widehat{c} = \frac{K}{\alpha}\Psi_q^{-1}\left(\gamma_q(c)\right). \tag{4.12}$$

When this happens, the $\widehat{c}$ given by (4.12) is asymptotically unbiased. However, the inversion Ψ_q^{-1} is not always allowed. There is a set of q's that can make Ψ_q invertible, which is defined as $\mathcal{Q}_\theta$ in Definition 2.3.1. The following proposition relates $\mathcal{Q}_\theta$ to $\gamma_q(c)$.

Proposition 4.2.2 *Let $0 < \delta < 1$ be a constant. Then, for any*

$$q \in \mathcal{Q}_\theta \stackrel{def}{=} \left\{ q \;\middle|\; 1 - \left(\frac{\delta}{2}\right)^{\frac{1}{KT}} \leq \Psi_q(\theta) \leq \left(\frac{\delta}{2}\right)^{\frac{1}{KT}} \right\}, \tag{4.13}$$

the random variable $\gamma_q(c)$ will not attain 0 or 1 with probability at least $1-\delta$, i.e.,

$$\mathbb{P}[0 < \gamma_q(c) < 1] > 1 - \delta.$$

In this case, the ML estimate $\widehat{c}$ is uniquely defined by (4.12).

Proof See Appendix C.7. ■

Before we proceed, let us look at some rough magnitude of the parameters in the following example.

Example 4 *Let the ground truth pixel value be $c = 0.5$. The sensor parameters are set as $T = 50$, $K = 4$, $\alpha = 300$. For a constant $\delta = 2 \times 10^{-4}$, the tolerance level is $\varepsilon = 1 - (\delta/2)^{1/KT} = 0.045$. Therefore, as long as $q \in \{q \,|\, 0.045 \leq \Psi_q(\theta) \leq 1 - 0.045\}$, which is the set $\{q \mid 28 \leq q \leq 48\}$, the probability that $\gamma_q(c)$ equals to 0 or 1 is upper bounded by $\delta = 2 \times 10^{-4}$.*

4.2.3 Gap between $\mathcal{Q}_\theta$ and q^*

The result in the previous subsection shows that as long as $q \in \mathcal{Q}_\theta$, the ML estimate is asymptotic unbiased. However, how is a $q \in \mathcal{Q}_\theta$ compared to the oracle threshold q^*? We answer this question in three parts.

First, does an asymptotically unbiased estimate maximize the SNR? The answer is no, because Proposition 4.1.4 states that if q^* is the optimal threshold, then

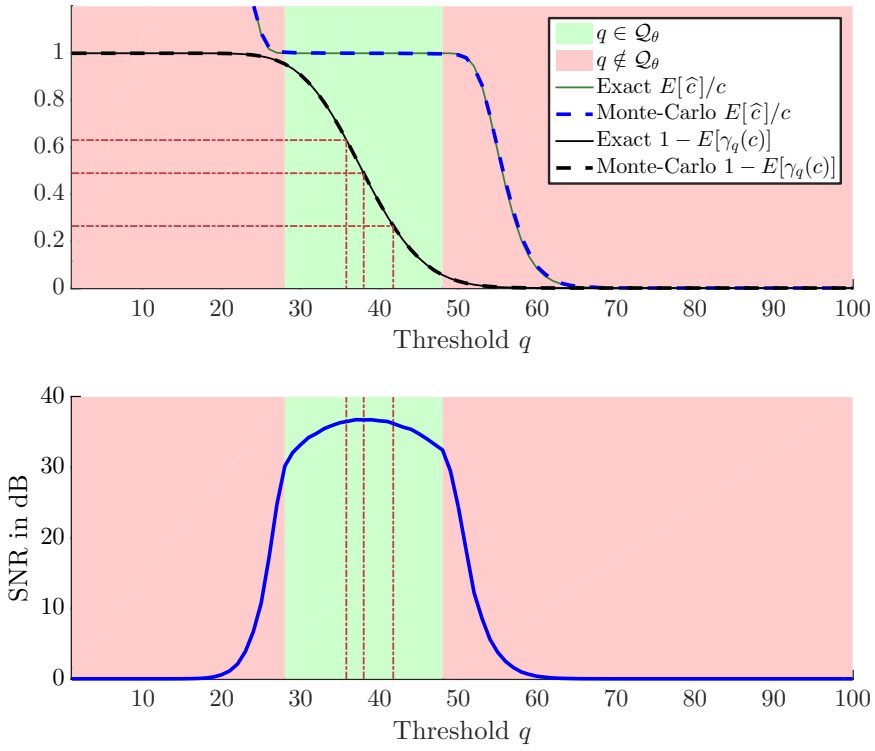


Fig. 4.2. Phase transition of the ML estimate and its relationship to the average bit density $1-\mathbb{E}[\gamma_q(c)]$. The red region is where it is impossible to recover c, whereas the green region is where we can have perfect recovery.

$\text{SNR}_{q^*}(c) \geq \text{SNR}_q(c)$ for any $q \neq q^*$. Therefore, moving from the exact optimal q^* to an asymptotically unbiased threshold is a relaxation of the optimality criteria.

If asymptotic unbiasedness is a relaxed optimality criteria, how much SNR drop will there be if we choose a $q \in \mathcal{Q}_\theta$ but not necessarily $q = q^*$? We show in Figure 4.2 the plot of a typical experiment with setup discussed in Example 4. As shown in the figure, the green zone is the set $\mathcal{Q}_\theta = \{q \mid 28 \leq q \leq 48\}$, or equivalently $\mathcal{Q}_\theta = \{q \,|\, 0.045 \leq \Psi_q(\theta) \leq 0.9955\}$. For any q in this $\mathcal{Q}_\theta$, the reconstruction has a SNR at least 30dB. If we further tighten $\mathcal{Q}_\theta$ so that $\mathcal{Q}_\theta = \{q \mid 35 \leq q \leq 42\}$, or equivalently $\mathcal{Q}_\theta = \{q \,|\, 0.25 \leq \Psi_q(\theta) \leq 0.6\}$, the SNR stays in the range $36.15\text{dB} \leq \text{SNR}_q(c) \leq 36.65\text{dB}$, which is reasonably narrow.

How tight should $\mathcal{Q}_\theta$ be? Ideally we want $\mathcal{Q}_\theta$ to be as tight as possible. But knowing the fact that the incomplete Gamma function has a rapid transition (See

the black line in Figure 4.2), $\mathcal{Q}_\theta$ can be much wider. In fact, we can choose $\mathcal{Q}_\theta$ such that $1-\gamma_q(c)$ stays close to 0.5, so that we are guaranteed to obtain a near optimal threshold. From an information theoretic point of view, $1-\gamma_q(c) \approx 0.5$ is where the bit density attains the maximum information — if q is too high then most bits become 0 whereas if q is too low then most bits become 1. It is maximum when q leads to 50% zeros and 50% ones. [2]

4.2.4 Phase Transition Phenomenon

We can now point out a very interesting phenomenon in Figure 4.2. In the upper plot of Figure 4.2 we show two sets of curves: blue curves (solid and dotted), and black curves (solid and dotted). The blue curves represent the ratio $\mathbb{E}[\widehat{c}]/c$, and the black curves represent the average bit density $1-\mathbb{E}[\gamma_q(c)]$. For both sets of curves, we use dotted lines to illustrate the Monte-Carlo simulation using 10,000 random samples, where each sample refers to a spatial-temporal block $\mathcal{B}_n$ containing $KT = 200$ binary bits. Notice that these dotted lines overlap exactly with their expectations, and hence (4.7)-(4.11) are valid.

Let us take a closer look at the blue curve $\mathbb{E}[\widehat{c}]/c$. Let $\mathcal{Q}_\theta = \{q \mid q_L \leq q \leq q_H\}$, where q_L and q_H are the smallest and the largest integers in $\mathcal{Q}_\theta$ respectively. There are three distinct phases:

- When $q < q_L$, the threshold is low and so most bits become 1. Therefore, $\gamma_c(q) \to 0$ and hence $\widehat{c} \to \infty$. Thus, $\mathbb{E}[\widehat{c}]/c \to \infty$ as q decreases.
- When $q > q_H$, the threshold high and so most bits become 0. Therefore, $\gamma_c(q) \to 1$ and hence $\widehat{c} \to 0$. Thus, $\mathbb{E}[\widehat{c}]/c \to 0$ as q increases.
- When $q_L \leq q \leq q_H$, the ML estimate $\widehat{c}$ is asymptotically unbiased. Therefore, $\mathbb{E}[\widehat{c}]/c = 1$.

[2] The exact optimal value of $1-\gamma_q(c)$ at q^* is slightly lower than 0.5 due to the nonlinearity of the Gamma function. See Supplementary Material for additional discussion.

Algorithm 1 Bisection Threshold Update Scheme

Initial thresholds q_A and q_B such that $1 - \gamma_{q_A} > 0.5$ and $1 - \gamma_{q_B} < 0.5$.

Compute $q_M = \lceil (q_A + q_B)/2 \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling operator.

while $|\gamma_{q_M} - 0.5| <$ `tol` **do**

If $\gamma_{q_M} < 0.5$, then set $q_A = q_M$. Else, set $q_B = q_A$.

Compute $q_M = \lceil (q_A + q_B)/2 \rceil$.

end while

return q_M

Essentially, Figure 4.2 demonstrates a phase transition behavior of the threshold. Such phase transition exists because Ψ_q is only invertible when $q \in \mathcal{Q}_\theta$.

4.2.5 Bisection Threshold Update Scheme

Now we present a practical threshold update scheme. As we discussed in Section IV.C, the oracle threshold q^* can be obtained when bit density $\gamma_q(c)$ is close to 0.5. Therefore, a practical procedure to determine q is to sweep through a range of q until the bit density reaches 0.5. To achieve this objective, we propose a bisection method illustrated in Figure 4.3 and Algorithm 1. Starting with initial thresholds q_A and q_B, we check whether the bit density satisfies $1 - \gamma_{q_A} > 0.5$ and $1 - \gamma_{q_B} < 0.5$. If this is the case, then we find a mid point $q_M = (q_A + q_B)/2$ and check whether $1 - \gamma_{q_M}$ is greater or less than 0.5. If $1 - \gamma_{q_M} > 0.5$, we replace q_A by q_M, otherwise we replace q_B by q_M. The process repeats until $1 - \gamma_{q_M}$ is sufficiently close to 0.5.

In our proposed threshold update scheme, we assume that the image has been partitioned into N blocks $\{\mathcal{B}_n \mid n = 0, \ldots, N-1\}$. Each $\mathcal{B}_n$ contains KT binary bits and is used to estimate one pixel value c_n. This setting results in N different thresholds, one for every pixel. To generalize the setting, it is also possible to allow multiple pixels to share a common threshold. Figure 4.4 shows an example. The advantage of sharing a threshold for multiple pixels is that circuits associated with

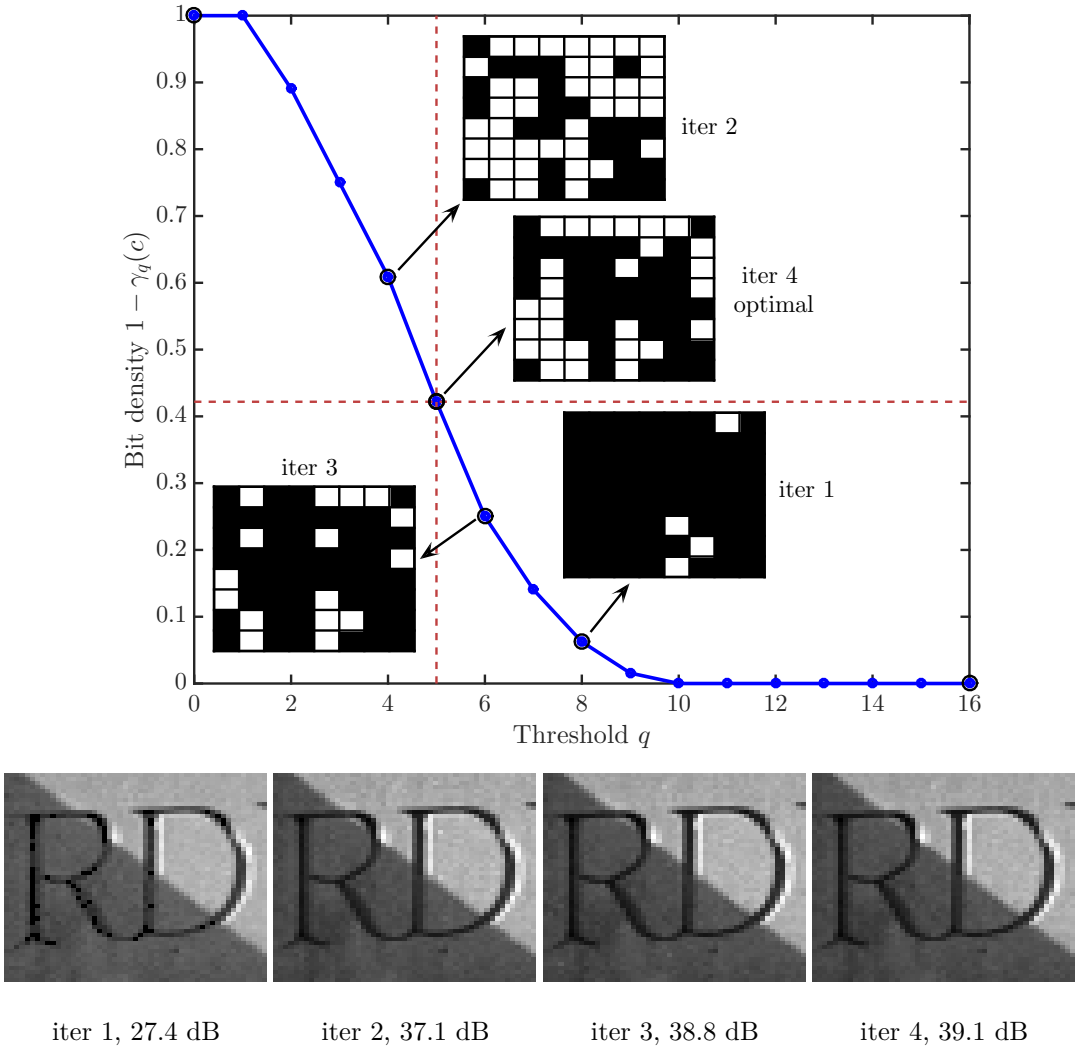


Fig. 4.3. The proposed bisection update scheme adjusts the threshold q such that the bit density $1-\gamma_q(c)$ approaches 0.5. The upper graph illustrates the bisection steps. Bottom row shows cropped patches from reconstructed images using threshold maps at different iterations and the PSNRs.

the sensor can be simplified. In terms of performance, since neighboring pixels are typically correlated, sharing the threshold causes little drop in the resulting SNR.

The price that the proposed bisection algorithm has to pay is the number of frames it requires to determine a good q. For every evaluation of γ_{q_M}, the sensor has to physically acquire one frame and compute the bit density in each of the N blocks.

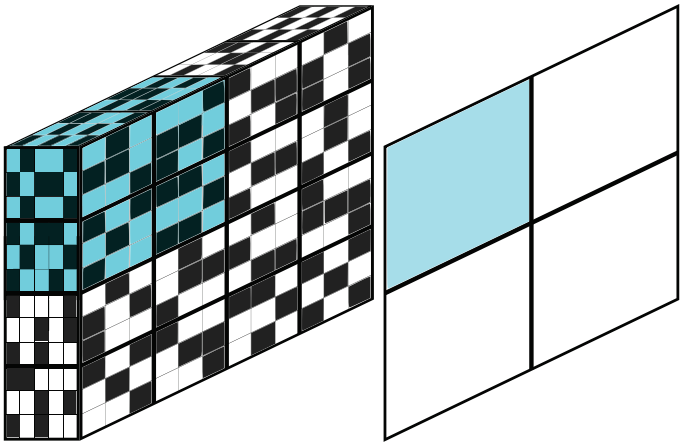


Fig. 4.4. Concept of shared thresholds. (Left) binary measurements, spatial oversampling $K = 3 \times 3$, Temporal oversampling $T = 5$. (Right) Threshold map, one threshold value is shared by 6×6 jots.

Therefore, the more bisection steps we need, the more frames that the sensor has to physically acquire. The rate of convergence of the proposed method and existing methods will be compared in the experimental results in Chapter 5.

4.2.6 Extension to High Dynamic Range

While QIS is a photon counting device, it is designed to count a few photons to keep the full-well capacity small, e.g. 20 photoelectrons as reported in [121]. Therefore, for practical imaging tasks, we need to extend the dynamic range for QIS.

There are two ways to enable dynamic range extension:

- Bright Scenes: Reduce Duty Cycle. In the signal processing block diagram shown in Figure 2.1, we can replace the constant α by a fraction as $\alpha\tau$, where $0 \leq \tau \leq 1$ determines the ratio between the actual integration time and the readout scan time. It can also be referred to the shutter duty cycle because the shutter is opened to collect photons during this proportion of time [122]. For very bright scenes, a low duty cycle will prevent QIS from saturating early.

- Dark Scenes: Multiple Measurements. For dark scenes, multiple measurements can be taken to ensure enough photons over the measurement period. This, however, is different from conventional HDR imaging. In conventional HDR imaging, the multiple shots are taken at different shutter speeds, e.g., 1/8192, 1/2048, 1/512, 1/128, 1/32, 1/8, 1/2 seconds [123], which is redundant. QIS's multiple shot functions more similar to burst photography [124]. The amount of acquisition time is significantly less than the conventional HDR imaging.

These two methods can be used for *any* threshold scheme, including ours and others. The benefit of using our proposed threshold scheme is that it supports a much wider dynamic range extension. In Figure 4.5, we illustrate the total dynamic range that can be covered using 4 multiple measurements at duty cycles $\tau = 1$, $\tau = 0.2$, $\tau = 0.04$, and $\tau = 0.008$. The maximum threshold level is $q_{\max} = 25$, and the minimum threshold level is $q_{\min} = 1$. It can be seen from the figure that with the optimal threshold q^*, the dynamic range is significantly more than the non-optimal ones. In particular, we observe a 16dB and a 54dB improvement compared to $q_{\min} = 1$ and $q_{\max} = 25$, respectively. Experimental results will be shown in Chapter 5

4.2.7 Hardware Consideration

Concerning the hardware implementation, we anticipate that future QIS will be equipped with per-pixel FPGAs to perform the proposed threshold update scheme. On-sensor FPGA is an actively developing technology. For example, MIT Lincoln Lab's digital focal plane array can achieve on-sensor image stabilization and edge detection [125] . For QIS threshold update, the complexity is low because we are only counting the number of ones in the bisection. More specifically, in order to perform the bisection, we only need K additions to compute $\sum_{k=0}^{K-1} b_{Kn+k,t}$; one comparison $\sum_{k=0}^{K-1} b_{Kn+k,t} \geq 0.5$; one addition and one multiplication (with a constant 0.5) to update the threshold $q_M = \lceil (q_A + q_B)/2 \rceil$. The dominating factor here is the K additions, which can be implemented efficiently by shifting bits in a buffer.

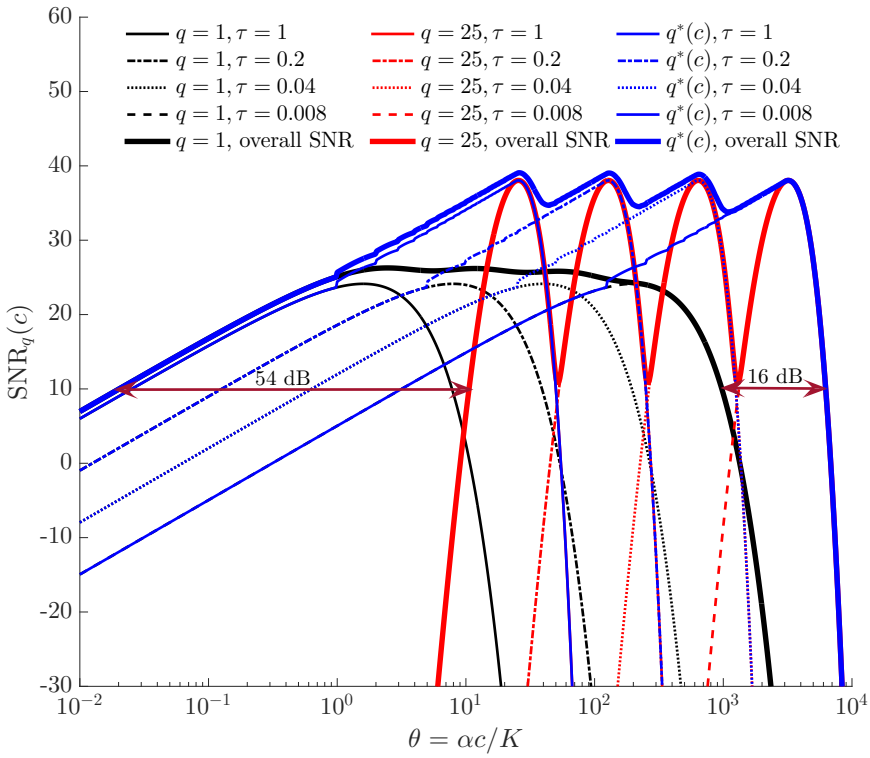


Fig. 4.5. SNR in dB vs. exposure θ for HDR imaging mode obtained by fusion of frames with shutter duty cycles $\tau \in \{1, 0.2, 0.04, 0.008\}$. Three scenarios are shown: constant threshold with $q = 1$ (black), $q = 25$ (red) and an optimal spatially varying threshold (blue).

We should also point out that the proposed bisection method can be flexibly adjusted spatially and temporally for different hardware configurations. For example, we can use a spatial-temporal window $4 \times 4 \times 1$ for low-resolution high-speed imaging, or $1 \times 1 \times 16$ for high-resolution low-speed imaging. This flexibility offers additional advantages of QIS over conventional CCD and CMOS cameras.

5. COLOR FILTER ARRAYS FOR QUANTA IMAGE SENSORS

This chapter presents an optimization-based framework to design color filter arrays for very small pixels. The new framework unifies several mainstream color filter array design frameworks by offering generality and flexibility. Compared to the existing frameworks which can only handle one or two design criteria, the new framework can simultaneously handle luminance gain, chrominance gain, cross-talk, anti-aliasing, manufacturability and orthogonality. Extensive experimental comparisons demonstrate the effectiveness and generality of the framework.

We start by providing a background and describing the notations of the imaging model in Section . Then, we present in Section the design criteria used for obtaining an efficient CFA. Afterwards, we solve the CFA design problem in Section 5.3. Finally, we present in Section 5.4 a universal demosaicking algorithm that can do demosaicking for any color filter array.

5.1 Background and Notations

The design of a robust CFA involves multiple objectives in terms of signal sensitivity, color aliasing, cross-talk, and manufacturability. To facilitate readers to understand the design framework, in this section we introduce a few notations and terminologies. We will start in Section 5.1.1 by describing the image formation using a CFA, then we discuss CFA in different domains in Sections 5.1.2 and 5.1.3. Afterwards, in Section 5.1.4, we define the optimization variables to simplify the design framework.

5.1.1 Color Image Formation

Consider a color image $\mathbf{im}_{rgb}$ of size $H \times W$. We denote the normalized light intensities in the red, green and blue channels for the (m-th,n-th) pixel of the color image as

$$\mathbf{im}_{rgb}(m,n) = \begin{bmatrix} \mathrm{im}_r(m,n) \\ \mathrm{im}_g(m,n) \\ \mathrm{im}_b(m,n) \end{bmatrix}, \tag{5.1}$$

where $m \in \{0,\ldots,H-1\}$, $n \in \{0,\ldots,W-1\}$.

Color Filtering: To obtain color, we place a color filter on top of each jot to collect light for one of the RGB colors. The CFA is a periodic pattern of the same resolution of $\mathbf{im}_{rgb}$, defined as

$$\boldsymbol{c}_{rgb}(m,n) = \begin{bmatrix} c_r(m,n) \\ c_g(m,n) \\ c_b(m,n) \end{bmatrix}, \tag{5.2}$$

where $c_r(m,n)$, $c_g(m,n)$, $c_b(m,n) \in [0,1]$ are the opacity rates for the red, green and blue pixels, respectively. For example, a red color filter is defined as $\boldsymbol{c}_{rgb}(m,n) = [1,0,0]^T$ as it only passes the red color. The light exposure on the QIS after passing through the CFA is denoted as $\theta(m,n)$, which is a linear combination of the tri-stimulus colors:

$$\begin{aligned} \theta(m,n) &= \alpha \boldsymbol{c}_{rgb}(m,n)^T \mathbf{im}_{rgb}(m,n) \\ &= \alpha \sum_{i\in\{r,g,b\}} c_i(m,n)\,\mathrm{im}_i(m,n). \end{aligned} \tag{5.3}$$

Here, α is a positive scalar representing the sensor gain factor.

Photon Arrival. The photon arrival is modeled as a Poisson process. Let $\boldsymbol{Y} \in \mathbb{N}^{HW}$ be a vector of non-negative random integers denoting the number of photons arriving at QIS jots according to the light exposure $\boldsymbol{\theta}$. Then, the probability of observing a photon count $Y_m = y_m$ is

$$\mathbb{P}(Y_m = y_m) = \frac{\theta_m^{y_m} e^{-\theta_m}}{y_m!}. \tag{5.4}$$

In this work, we assume *single-bit QIS* [26] that quantizes the photon counts by QIS jots to a binary values $\boldsymbol{B} \in \{0,1\}^{HW}$ with $B_m = 1$ if $Y_m \geq q$ and $B_m = 0$ if $Y_m < q$, where $q > 0$ is a threshold. The probability of observing $B_m = b_m$ is

$$\mathbb{P}(B_m = b_m) = \Psi_q(\theta_m)^{1-b_m} \left(1 - \Psi_q(\theta_m)\right)^{b_m}, \tag{5.5}$$

where $\Psi_q(.)$ is the incomplete Gamma function [69].

Temporal Oversampling. With frame rates that reach 1000 fps, QIS is able to catch the scene movement by taking T temporal samplings for the same scene. This allows us to utilize multiple independent measurements over time to improve the statistics and decrease noise. Hence, for every jot with light exposure θ_m, we have a set of T independent binary measurements $\mathcal{B}_m = \{b_{m,0}, \ldots, b_{m,T-1}\}$.

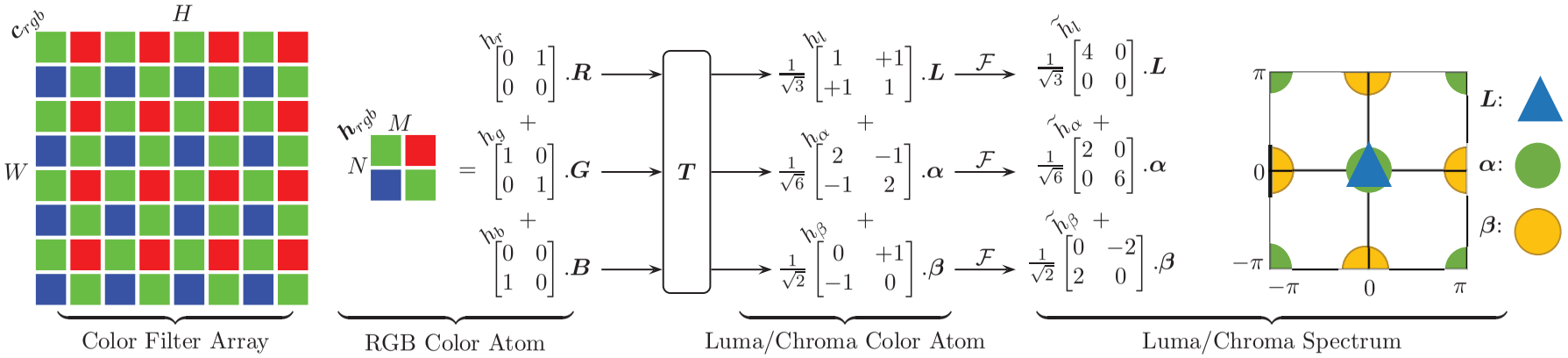


Fig. 5.1. Our terminology illustrated on the Bayer CFA example. The building unit of a CFA is a color atom. A transformation T is applied to the color atom to transform it from the canonical RGB color space to a luma/chroma color space to simplify the design process. Foruier transform is applied afterwards to obtain the color atom spectrum in the luma/chroma space.

5.1.2 Color Filter Array Analysis in Different Color Spaces

Since the CFA $\boldsymbol{c}_{rgb}(m,n)$ is a periodic pattern, it is sufficient to use a *color atom* as the optimization variable when designing the CFA. The color atom takes the form

$$\boldsymbol{h}_{rgb}(m,n) = \begin{bmatrix} h_r(m,n) \\ h_g(m,n) \\ h_b(m,n) \end{bmatrix}, \tag{5.6}$$

where each of h_r, h_g and h_b is an $M \times N$ array. For instance, the GRBG Bayer pattern has the following color atom (when $M = N = 2$):

$$h_r = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},\ h_g = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, h_b = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$

because the Bayer pattern has one red pixel and one blue pixel located at two opposite diagonals, and two green pixels located in the remaining two positions. Figure 5.1 illustrates the idea.

While the primal RGB color is common for making the CFA, it would be more convenient if the colors are *decorrelated*. To this end, we change the image representation from the canonical RGB basis to an orthornormal basis using a transformation matrix [95, 126]:

$$\boldsymbol{T} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1/\sqrt{6} & 2/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix}. \tag{5.7}$$

This transformation maps an RGB image $\mathbf{im}_{rgb}$ to an image $\mathbf{im}_{l\alpha\beta} = [\mathrm{im}_l, \mathrm{im}_\alpha, \mathrm{im}_\beta]^T$ as follows (we drop the spatial indices (m,n) for simplicity)

$$\mathbf{im}_{l\alpha\beta} = \begin{bmatrix} \mathrm{im}_l \\ \mathrm{im}_\alpha \\ \mathrm{im}_\beta \end{bmatrix} = \boldsymbol{T} \begin{bmatrix} \mathrm{im}_r \\ \mathrm{im}_g \\ \mathrm{im}_b \end{bmatrix} = \begin{bmatrix} (\mathrm{im}_r + \mathrm{im}_g + \mathrm{im}_b)/\sqrt{3} \\ (-\mathrm{im}_r + 2\mathrm{im}_g + \mathrm{im}_b)/\sqrt{6} \\ (\mathrm{im}_r - \mathrm{im}_b)/\sqrt{2} \end{bmatrix},$$

where im_l is a luminance (luma) component that contains the high frequency components such as edges and textures, whereas im_α and im_β are chrominance (chroma) components that carry the color information.

Since $\boldsymbol{T}$ is orthonormal (i.e., $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$), we can rewrite the sampling process in (B.8) in the luma/chroma space:

$$\begin{aligned}\theta(m,n) &= \alpha \boldsymbol{c}_{rgb}(m,n)^T\, \boldsymbol{T}^T\boldsymbol{T}\, \mathbf{im}_{rgb}(m,n) \\ &= \alpha \boldsymbol{c}_{l\alpha\beta}(m,n)^T\, \mathbf{im}_{l\alpha\beta}(m,n) \\ &= \alpha \sum_{i\in\{l,\alpha,\beta\}} c_i(m,n)\, \mathrm{im}_i(m,n), \end{aligned} \tag{5.8}$$

where $c_l(m,n)$, $c_\alpha(m,n)$ and $c_\beta(m,n)$ are the luma/chroma representation of the CFA, with

$$\boldsymbol{c}_{l\alpha\beta}(m,n) = \boldsymbol{T}\, \boldsymbol{c}_{rgb}(m,n). \tag{5.9}$$

The luma/chroma representation of the CFA has a corresponding color atom $h_l(m,n)$, $h_\alpha(m,n)$ and $h_\beta(m,n)$. For instance, the luma/chroma color atom of the GRBG Bayer pattern is

$$\boldsymbol{h}_{l\alpha\beta}(m,n) = \begin{bmatrix} h_l(m,n) \\ h_\alpha(m,n) \\ h_\beta(m,n) \end{bmatrix}, \tag{5.10}$$

where the individual components are

$$h_l = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, h_\alpha = \frac{1}{\sqrt{6}}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, h_\beta = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Remark 4 *In principle, there are are infinite choices for the luma/chroma basis* $\boldsymbol{T}$*. We choose the one in* (5.7) *because it makes the components of natural images statistically independent in the first order approximation.*

5.1.3 Color Filter Array in Fourier Space

When analyzing the aliasing effects of the CFAs, we need to transform the color atom into the Fourier domain. For simplicity, we represent the Fourier transform of a signal by putting a tilde on top of the symbol, e.g., $h \overset{\mathcal{F}}{\rightarrow} \widetilde{h}$. The 2D discrete Fourier transform (DFT) of the i-th color atom is

$$\widetilde{h}_i(u,v) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} h_i(m,n) e^{-j2\pi\left(\frac{mu}{M}+\frac{nv}{N}\right)} \tag{5.11}$$

$$\widetilde{h}_i = \begin{bmatrix} \delta_{00} & \delta_{01} & \delta_{02} \\ \delta_{10} & \delta_{11} & \delta_{12} \\ \delta_{20} & \delta_{21} & \delta_{22} \end{bmatrix} \overset{\text{Vec}}{\rightarrow} \widetilde{\boldsymbol{h}}_i = \begin{bmatrix} \delta_{00} \\ \delta_{10} \\ \delta_{20} \\ \delta_{01} \\ \delta_{11} \\ \delta_{21} \\ \delta_{02} \\ \delta_{12} \\ \delta_{22} \end{bmatrix}$$

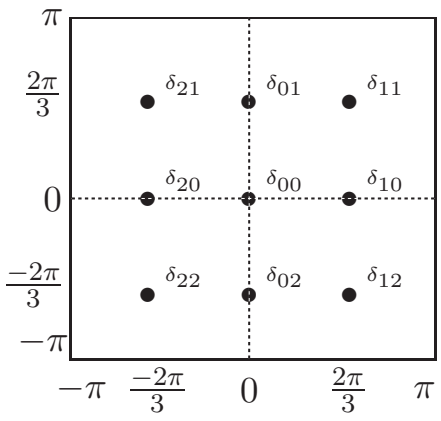


Fig. 5.2. The Fourier representation of an arbitrary 3×3 color atom i. From left to right: The atom representation, the vector representation and the 2D frequency plane representation. Notice that the frequency plane is divided into 9 regions of size $2\pi/3$, and the spectrum comprises pure sinusoids placed at $(\frac{2\pi u}{3}, \frac{2\pi v}{3})$ $\forall u, v \in \{0, 1, 2\}$.

where $u \in \{0, \ldots, M-1\}$, $v \in \{0, \ldots, N-1\}$.

For example, the discrete Fourier transform of the luma/chroma color atoms in (5.10) are

$$\widetilde{h}_l = \frac{1}{\sqrt{3}} \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}, \widetilde{h}_\alpha = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}, \widetilde{h}_\beta = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}.$$

Here, we observe that the Fourier transform of the color atom has the same size as the original color atom. The luminance channel has only one baseband components at $(0, 0)$, whereas the α chrominance channel has one baseband component and a component at (π, π). The β chrominance channel has two components at $(0, \pi)$ and $(\pi, 0)$. Figure 5.2 illustrates how these frequency locations are identified from a 3×3 color atom.

While the Fourier transform of the color atom is useful, for demosaicing we also need to analyze the spectrum of the entire CFA. As shown in by Hao et al. [86], the

Fourier transform of the entire CFA can be written in terms of the Fourier transform of the color atoms:

$$\widetilde{c}_i(\boldsymbol{\omega}) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \widetilde{h}_i(u,v)\delta\left(\boldsymbol{\omega} - \boldsymbol{\omega}(u,v)\right), \tag{5.12}$$

where $i \in \{l, \alpha, \beta\}$, $\boldsymbol{\omega}$ is the 2D angular frequency, and

$$\boldsymbol{\omega}(u,v) = \left(\frac{2\pi u}{M}, \frac{2\pi v}{N}\right) \; \forall \; \begin{matrix} u \in \{0,\ldots,M-1\} \\ v \in \{0,\ldots,N-1\} \end{matrix}. \tag{5.13}$$

is the (u-th,v-th) 2D angular frequency. It is worth noting that the Fourier transform of the CFA comprises pure sinusoids of amplitudes $\widetilde{h}_i(u,v)$. These sinusoids are placed at MN discrete 2D frequencies $\boldsymbol{\omega}(u,v)$ that divide the 2D frequency plane $[-\pi,\pi] \times [-\pi,\pi]$ into MN equal regions. Therefore, the spectrum of the mosaicked image $\widetilde{\theta}(\boldsymbol{\omega})$ can be written as

$$\begin{aligned} \widetilde{\theta}(\boldsymbol{\omega}) &= \mathcal{F}\left(\sum_{i\in\{l,\alpha,\beta\}} c_i \, \mathrm{im}_i\right) = \sum_{i\in\{l,\alpha,\beta\}} \widetilde{c}_i(\boldsymbol{\omega}) \circledast \widetilde{\mathrm{im}}_i(\boldsymbol{\omega}) \\ &= \sum_{i\in\{l,\alpha,\beta\}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \widetilde{h}_i(u,v) \, \widetilde{\mathrm{im}}_i(\boldsymbol{\omega} - \boldsymbol{\omega}(u,v)), \end{aligned} \tag{5.14}$$

where $\circledast$ is the standard 2D convolution operator. Having the spectrum of the mosaicked image $\widetilde{\theta}(\boldsymbol{\omega})$, we can now discuss the optimization variables in our problem.

5.1.4 Design Variables

To formulate the CFA design problem as an optimization problem, we define the following variables. We denote $\boldsymbol{h}_r$, $\boldsymbol{h}_g$ and $\boldsymbol{h}_b$ the vectorized representations of the red, green and blue color atoms, respectively. To ensure physical realizability, we require $\boldsymbol{h}_r, \boldsymbol{h}_g, \boldsymbol{h}_b \in [0,1]^{K\times 1}$, where $K \stackrel{\text{def}}{=} MN$, and we stack all design variables into one long vector

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{h}_r \\ \boldsymbol{h}_g \\ \boldsymbol{h}_b \end{bmatrix} \in \mathbb{R}^{3K\times 1}.$$

The design variable $\boldsymbol{x}$ is related to the vectorized RGB and luma/chroma color atoms as

$$\begin{bmatrix} \boldsymbol{h}_r \\ \boldsymbol{h}_g \\ \boldsymbol{h}_b \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_r \\ \boldsymbol{Z}_g \\ \boldsymbol{Z}_b \end{bmatrix} \boldsymbol{x} \quad \text{and} \quad \begin{bmatrix} \boldsymbol{h}_l \\ \boldsymbol{h}_\alpha \\ \boldsymbol{h}_\beta \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_l \\ \boldsymbol{Z}_\alpha \\ \boldsymbol{Z}_\beta \end{bmatrix} \boldsymbol{x},$$

where the $\boldsymbol{Z}$ matrices are defined by (5.7) as

$$\begin{aligned} \boldsymbol{Z}_r &= [\boldsymbol{I}, \boldsymbol{0}, \boldsymbol{0}] & \boldsymbol{Z}_l &= [\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{I}]/\sqrt{3}, \\ \boldsymbol{Z}_g &= [\boldsymbol{0}, \boldsymbol{I}, \boldsymbol{0}] \quad \text{and} & \boldsymbol{Z}_\alpha &= [-\boldsymbol{I}, 2\boldsymbol{I}, -\boldsymbol{I}]/\sqrt{6}, \\ \boldsymbol{Z}_b &= [\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{I}] & \boldsymbol{Z}_\beta &= [\boldsymbol{I}, \boldsymbol{0}, -\boldsymbol{I}]/\sqrt{2}. \end{aligned}$$

Given the design variable $\boldsymbol{x}$, we also need to analyze its spectrum. We write the 2D Fourier transform equation (5.11) as a matrix-vector product by using the Fourier transform matrix $\boldsymbol{F} \in \mathcal{C}^{K \times K}$. Hence, the vectorized spectra of the luma/chroma color atoms can be written in terms of $\boldsymbol{x}$ as

$$\widetilde{\boldsymbol{h}}_i = \boldsymbol{F}\boldsymbol{h}_i = \boldsymbol{F}\boldsymbol{Z}_i\boldsymbol{x}, \; i \in \{l, \alpha, \beta\}. \tag{5.15}$$

where $\widetilde{\boldsymbol{h}}_i \in \mathcal{C}^{K \times 1}$, for $i \in \{l, \alpha, \beta\}$. The relation between the matrix and the vector forms of the Fourier transform is:

$$\widetilde{h}_i(u, v) = \text{vec}^{-1}(\widetilde{\boldsymbol{h}}_i) \tag{5.16}$$

where $\widetilde{h}_i(u, v)$ is the Fourier coefficient.

5.2 Design Criteria

We now present the design criteria. Our criteria unify the three major approaches in the literature: (i) Sensitivity of luma/chroma channels to noise by Condat [95]; (ii) Aliasing between different color components in the frequency domain by Hirakawa and Wolfe [83]; (iii) Crosstalk between neighboring pixels in the spatial domain by Anzagira and Fossum [93]. Note that the first two criteria were developed for CMOS, whereas the third criterion was developed for QIS. The proposed framework integrates

all these criteria into a unified formulation. Table 5.1 summarizes the difference between this paper and the previous works.

In the following subsections, we present the design criteria and express them in terms of matrix-vectors for the optimization framework in Section IV.

Table 5.1.

CFA Design Criteria

Criterion	Purpose	Regular Pixels				QIS	
		[83]	[95]	[86]	[98]	[93]	Ours
Proposition 5.2.1	To minimize noise power after linear demosaicking	×	✓	×	×	×	✓
Proposition 5.2.2	To simplify denoising of luminance channel	✓	✓	✓	✓	✓	✓
Proposition 5.2.3	To simplify denoising of chrominance channel	✓	✓	×	×	×	✓
Proposition 5.2.4	To maximize spatial resolution	✓	✓	✓	×	×	✓
Proposition 5.2.5	To mitigate cross-talk	×	×	×	✓	✓	✓
Definition 5.2.4	Enforce total orthogonality	✓	×	✓	×	×	✓
Definition 5.2.4	Enforce quadrature orthogonality	×	✓	×	×	×	✓

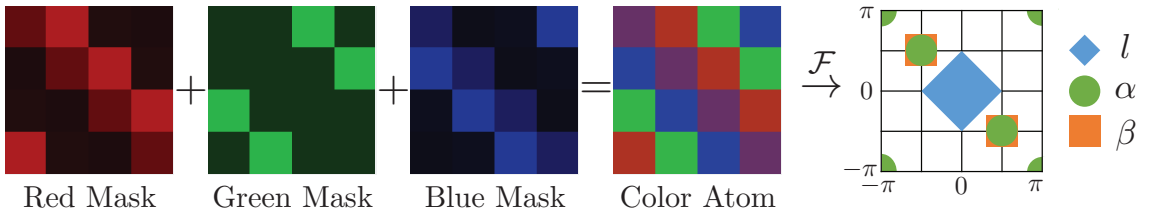


Fig. 5.3. A 4×4 CFA generated by our design framework. Luminance sensitivity γ_l and chrominance sensitivity γ_c are maximized to improve robustness to noise (Section III-A). No chrominance components (α or β) are modulated on the vertical and horizontal frequency axes (Section III-B) to mitigate aliasing with the luminance component l. The total variation of the red, green and blue masks is upper-bounded by $\mathrm{TV}_{\max}$ to mitigate crosstalk (Section III-C).

5.2.1 Luminance and Chrominance Sensitivity

Definition 5.2.1 *The luminance sensitivity γ_l and the chrominance sensitivity γ_c of a CFA with color atom $\{\boldsymbol{h}_l, \boldsymbol{h}_\alpha, \boldsymbol{h}_\beta\}$ of size $M \times N$ are defined as*

$$\gamma_l \stackrel{def}{=} \frac{1}{K}||\widetilde{\boldsymbol{h}}_l||_2, \text{ and } \gamma_c \stackrel{def}{=} \frac{1}{K} min\left(||\widetilde{\boldsymbol{h}}_\alpha||_2, ||\widetilde{\boldsymbol{h}}_\beta||_2\right). \tag{5.17}$$

where $K = MN$ is a normalization factor.

Intuitively, the luminance and chrominance sensitivity are measures of the signal power that can be transmitted through the color filter. A more transparent color filter allows more light to pass through, and hence the signal power is higher. This is reflected by the magnitudes $\|\widetilde{\boldsymbol{h}}_i\|_2$ for $i \in \{l, \alpha, \beta\}$, which according to Parseval's Theorem they are equivalent to $\|\boldsymbol{h}_i\|_2$.

The following proposition shows how can we compute γ_l and γ_c in terms of the optimization vector $\boldsymbol{x}$.

Proposition 5.2.1 *For a CFA with color atoms represented by the vector $\boldsymbol{x}$, the luminance and chrominance sensitivity can be calculated as*

$$\begin{aligned} \gamma_l(\boldsymbol{x}) &= \frac{1}{K}\mathbf{1}^T \boldsymbol{Z}_l \boldsymbol{x} = \boldsymbol{b}^T \boldsymbol{x} \\ \gamma_c(\boldsymbol{x}) &= \min\left(\sqrt{\boldsymbol{x}^T \boldsymbol{Q}_\alpha \boldsymbol{x}}, \sqrt{\boldsymbol{x}^T \boldsymbol{Q}_\beta \boldsymbol{x}}\right), \end{aligned} \tag{5.18}$$

where $\boldsymbol{b} = \frac{1}{K}\boldsymbol{Z}_l^T\mathbf{1}$, $\boldsymbol{Q}_\alpha = \boldsymbol{Z}_\alpha^T\boldsymbol{Z}_\alpha$ *and* $\boldsymbol{Q}_\beta = \boldsymbol{Z}_\beta^T\boldsymbol{Z}_\beta$.

Proof See Appendix A. ■

The luminance sensitivity and the chrominance sensitivity cannot be arbitrarily chosen. One practical consideration is to ensure uniform noise power across the luma channel so that the denoising procedure can be simplified (because the noise will be i.i.d.). Thus, the luminance color atom h_l should be constant, i.e., $h_l(m,n) = c, \forall m, n$, where c is a positive constant. Taking Fourier transform, this means that $\widetilde{h}_l$ comprises only one impulse at baseband $\widetilde{h}_l(0,0)$, and no impulses at all other frequencies. In vector form, we need

$$\widetilde{\boldsymbol{h}}_l - \mathrm{diag}(\boldsymbol{e}_1)\widetilde{\boldsymbol{h}}_l = \mathbf{0}, \tag{5.19}$$

where $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^T$ is the standard basis. Putting in terms of the optimization variable $\boldsymbol{x}$, we have a constraint.

Proposition 5.2.2 (Uniform Luminance Constraint) *If a CFA has a uniform luminance sensitivity, then* $\boldsymbol{x}$ *needs to satisfy*

$$(\boldsymbol{I} - \mathit{diag}(\boldsymbol{e}_1))\boldsymbol{F}\boldsymbol{Z}_l\,\boldsymbol{x} = \mathbf{0}. \tag{5.20}$$

Proof Using (5.15), substitute $\widetilde{\boldsymbol{h}}_l = \boldsymbol{F}\boldsymbol{Z}_l\boldsymbol{x}$ into (5.19). ■

Similarly, we can impose a constraint to the chrominance channel. For chrominance, we require that the color filter passes the same amount of red, green, and blue so that the primary colors have uniform sensitivity [89, 95]. This leads to

$$\sum_{m,n} h_r(m,n) = \sum_{m,n} h_g(m,n) = \sum_{m,n} h_b(m,n)$$

Putting into vector form, we have the following constraint.

Proposition 5.2.3 (Uniform Chrominance Constraint) *If a CFA has a uniform chrominance sensitivity, then* $\boldsymbol{x}$ *needs to satisfy*

$$\begin{bmatrix} \boldsymbol{u}_R^T - \boldsymbol{u}_G^T \\ \boldsymbol{u}_R^T - \boldsymbol{u}_B^T \\ \boldsymbol{u}_G^T - \boldsymbol{u}_B^T \end{bmatrix} \boldsymbol{x} = \boldsymbol{U}\boldsymbol{x} = \mathbf{0}, \tag{5.21}$$

where $\boldsymbol{u}_R \stackrel{def}{=} [\mathbf{1}^T, \mathbf{0}^T, \mathbf{0}^T]$, $\boldsymbol{u}_R \stackrel{def}{=} [\mathbf{0}^T, \mathbf{1}^T, \mathbf{0}^T]$, *and* $\boldsymbol{u}_R \stackrel{def}{=} [\mathbf{0}^T, \mathbf{0}^T, \mathbf{1}^T]$.

5.2.2 Anti-Aliasing

In the frequency domain, the luminance controls the baseband whereas the chrominance controls the sideband of the spectrum. To minimize spectral interference, aka aliasing, it is necessary to modulate the chrominance as far as possible from the baseband. However, the luminance has approximately a diamond shape since it has large frequency components at $(\pm\pi, 0)$ and $(0, \pm\pi)$. Therefore, to mitigate aliasing, we should avoid modulating the chrominance on vertical and horizontal axes. Figure 5.5 shows a 4×4 CFA obtained by our framework. In this example, no chrominance components are modulated on the vertical and horizontal frequency axes.

The anti-aliasing requirement can be formulated as forcing the Fourier coefficients of the chrominance color atoms at $(\pm\pi, v)$ and $(u, \pm\pi)$ to zero for all u and v. This translates to chrominance color atom whose first column and first row are zeroed out (See Figure 5.2). In terms of our design variable $\boldsymbol{x}$, we require the following constraint.

Proposition 5.2.4 (Anti-aliasing Constraint) *The chrominance in the vertical and horizontal directions must be set to 0. Hence,* $\boldsymbol{x}$ *must satisfy*

$$\begin{bmatrix} \boldsymbol{W}_\alpha \\ \boldsymbol{W}_\beta \end{bmatrix} \boldsymbol{x} = \boldsymbol{W}\boldsymbol{x} = \mathbf{0} \tag{5.22}$$

where $\boldsymbol{W}_\alpha$ *and* $\boldsymbol{W}_\beta$ *are the matrices formed by choosing the rows in* $\boldsymbol{F}\boldsymbol{Z}_\alpha$ *and* $\boldsymbol{F}\boldsymbol{Z}_\beta$*, respectively, that correspond to the vertical and horizontal frequency components, i.e., rows in the set* $\{0, 1, \ldots, M-1\} \cup \{M, 2M, \ldots, (N-1)M\}$.

To quantify the amount of aliasing for every CFA, we define the following aliasing criterion.

Definition 5.2.2 *For a CFA, aliasing between luminance and chrominance channels is measured by*

$$J_l \stackrel{def}{=} \frac{1}{HW} \int_{[-\pi,\pi)^2} \left(S_l(\boldsymbol{\omega})S_\alpha(\boldsymbol{\omega}) + S_l(\boldsymbol{\omega})S_\beta(\boldsymbol{\omega})\right) d\boldsymbol{\omega}, \tag{5.23}$$

where S_l, S_α and S_β denote the power spectral density of the luminance channel im_l and the two chrominance channels im_α and im_β, respectively.

5.2.3 Crosstalk

Crosstalk is caused by the leakage of electrical and optical charge from a pixel to its adjacent pixels [93, 127]. Crosstalk leads to color de-saturation. To model crosstalk, we follow [93] by defining three scalars α_r, α_g, and α_b representing the proportion of leaked charges to neighboring pixels. These three scalars then form a *crosstalk kernel*,

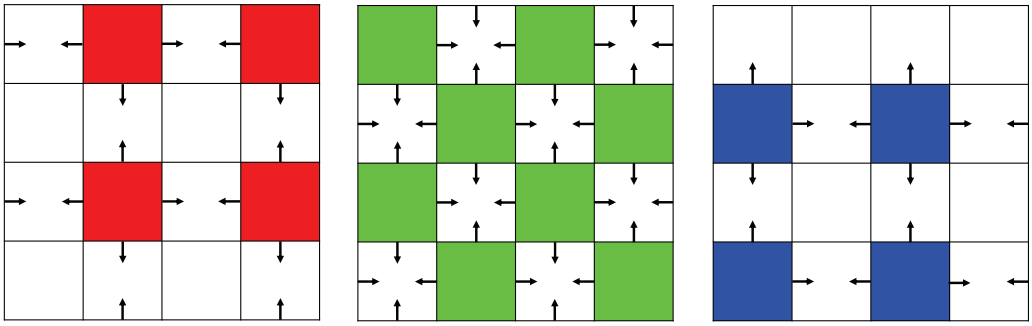
$$g_i = \begin{bmatrix} 0 & \alpha_i/4 & 0 \\ \alpha_i/4 & 1-\alpha_i & \alpha_i/4 \\ 0 & \alpha_i/4 & 0 \end{bmatrix}, \; i \in \{r, g, b\}, \tag{5.24}$$

which can be considered as a spatial lowpass filter of the mosaicked image. Applying the crosstalk kernel to the CFA is equivalent to a spatially invariant convolution

$$h_i^{\text{ctk}} = g_i \circledast h_i, \; i \in \{r, g, b\},$$

where h_i^{ctk} denotes the effective CFA in the presence of crosstalk.

The effect of crosstalk is more severe when the adjacent colors are different. For example, in Figure 5.4, the red and blue pixels are surround by 8 neighbors of different colors and the green pixels are surrounded by 4 neighbors of different colors. This is equivalent to saying that there is a red pixel having a value 1 and is surrounded by pixels having the value 0. Using similar argument, we can see that if the color atoms have more rapid changes of colors, then the resulting CFA is more susceptible to crosstalk.



(a) Red Channel (b) Green Channel (c) Blue Channel

Fig. 5.4. Crosstalk in Bayer Color Atom. Each color pixel leak some of its charge to its horizontal and vertical neighbors. Amount of leakage is parametrized by the positive scalars α_r, α_g and α_b.

We propose to quantify the variation of the color atoms (and hence crosstalk) is by means of measuring the total variation of the color atom. The total variation is a proxy of the complexity of the color filter array. A color filter array with high total variation means a more complicated pattern and so it is more susceptible to crosstalk. Our total variation is defined as follows.

Definition 5.2.3 (Total Variation) *For a CFA defined by the color atoms $\boldsymbol{h}_r$, $\boldsymbol{h}_g$ and $\boldsymbol{h}_b$, the weighted total variation is defined as*

$$TV(\boldsymbol{x}) \stackrel{def}{=} \sum_{i\in\{r,g,b\}} \alpha_i \|\boldsymbol{D}\boldsymbol{h}_i\|_1 = \sum_{i\in\{r,g,b\}} \alpha_i \|\boldsymbol{D}\boldsymbol{Z}_r\boldsymbol{x}\|_1 \tag{5.25}$$

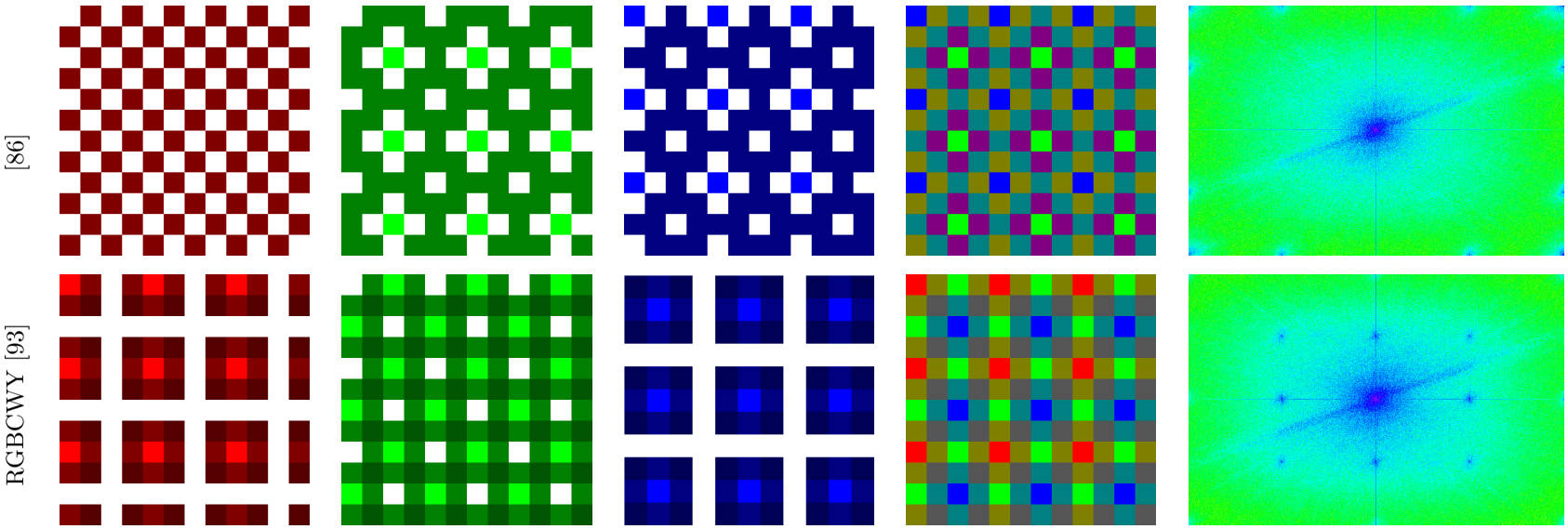
where $\boldsymbol{D} \stackrel{def}{=} [\boldsymbol{D}_x, \boldsymbol{D}_y]^T$ is an operator that computes the vertical and horizontal derivatives with circular boundary conditions, and α_i is the leakage factor defined in the crosstalk kernel in (B.14).

To control the amount of variations in the CFA (so that we can limit the amount of crosstalk), we upper bound the total variation by a scalar $\text{TV}_{\max}$. This leads to the following constraint.

Proposition 5.2.5 (Crosstalk Constraint) *The crosstalk is limited by upper-bounding the total variation metric* $\mathrm{TV}(\boldsymbol{x})$*:*

$$TV(\boldsymbol{x}) = \sum_{i \in \{r,g,b\}} \alpha_i ||\boldsymbol{D}\boldsymbol{Z}_r \boldsymbol{x}||_1 \leq TV_{\max}. \tag{5.26}$$

Figure 5.5 shows two CFAs proposed in literature. The first one, proposed in [86] is more robust to aliasing than the second one proposed in [93]. This is because the chrominance channels are modulated at high frequencies which are far from baseband luminance. However, [93] is more robust to crosstalk than [86] because the color atom have less variation in colors. We can also see this in the total variation values (0.413 compared to 0.263). This trade-off constitutes a gap in literature: Color filter designs can improve robustness of either aliasing or crosstalk, but not for both. Our proposed design framework allows us to optimize them simultaneously.



(a) Red Mask (b) Green Mask (c) Blue Mask (d) CFA Atom (e) Spectrum

Fig. 5.5. Examples of two CFAs. (Top row) Proposed in [86], this array has good aliasing properties, where chrominance channels are placed far away from luminance channel, but it has bad crosstalk properties: $\mathrm{TV}(\boldsymbol{x}) = 0.413$. (Bottom row) Proposed in [93], this array has good crosstalk properties $\mathrm{TV}(\boldsymbol{x}) = 0.263$, but it has bad aliasing properties.

5.2.4 Condition Number

When designing a color filter array, one should also be aware of the simplicity of the demosaicking algorithm. Since the luminance/chrominance transformation, color filtering and crosstalk are all linear processes, we can represent them by an overall color acquisition matrix $\boldsymbol{A}$. To demosaic the image, in principle we need to invert the $\boldsymbol{A}$ matrix. To avoid the amplification of the estimation error of luminance and chrominance channels, the condition number of $\boldsymbol{A}$ should be minimized for numerical stability. This metric was discussed in [86], but the authors considered the condition number of the luminance/chrominance transformation matrix $\boldsymbol{T}$ only. In our case, we generalize this metric by taking the color filtering and crosstalk into account as well.

To represent the image acquisition in frequency domain as a linear process, we assume the crosstalk kernels for red, green and blue pixels are the same $g_r = g_g = g_b$. Define the following frequency domain variables:

$$\widetilde{\mathbf{im}}_{rgb} = \begin{bmatrix} \widetilde{\mathbf{im}}_r^T \\ \widetilde{\mathbf{im}}_g^T \\ \widetilde{\mathbf{im}}_b^T \end{bmatrix}, \widetilde{H} = [\widetilde{\boldsymbol{h}}_l, \widetilde{\boldsymbol{h}}_\alpha, \widetilde{\boldsymbol{h}}_\beta], \text{ and } \widetilde{\boldsymbol{G}} = \text{diag}(\widetilde{\boldsymbol{g}}) \tag{5.27}$$

where $\widetilde{\boldsymbol{g}} \overset{\mathcal{F}}{\leftarrow} \boldsymbol{g}$ is the vectorized version of the $M \times N$ discrete Fourier transform of the crosstalk kernel g. Hence, the mosaicked image $\widetilde{\theta}$ can be written as

$$\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{G}}\widetilde{\boldsymbol{H}}\boldsymbol{T}\widetilde{\mathbf{im}}_{rgb} = \boldsymbol{A}\widetilde{\mathbf{im}}_{rgb} \tag{5.28}$$

where we define the color acquisition matrix as $\boldsymbol{A} \overset{\text{def}}{=} \widetilde{\boldsymbol{G}}\widetilde{\boldsymbol{H}}\boldsymbol{T}$. Denote by $\kappa(\boldsymbol{A})$ the condition number of $\boldsymbol{A}$, i.e.,

$$\kappa(\boldsymbol{A}) = \text{cond}(\boldsymbol{A}) \in [1, \infty] \tag{5.29}$$

Low values of $\kappa(\boldsymbol{A})$ imply stable demosaicking process that involves mild amplification of estimation errors in the luminance and chrominance components.

5.2.5 Orthogonality of Chrominance Channels

When designing a CFA, one should take into consideration of the complexity of the demosaicking process. Recall (5.14) where we show that

$$\widetilde{\theta}(\boldsymbol{\omega}) = \sum_{i\in\{l,\alpha,\beta\}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \widetilde{h}_i(u,v)\, \widetilde{\mathrm{im}}_i(\boldsymbol{\omega} - \boldsymbol{\omega}(u,v)).$$

This is a modulation of the signal by a modulating frequency $\boldsymbol{\omega}(u,v)$. Therefore, to reconstruct the signal, one approach is to demodulate by shifting the channels to the baseband by multiplying pure sinusoids and then applying a lowpass filter [95]. Demodulation can be done efficiently if there is orthogonality between the channels. Following the literature, our optimization takes into account of two forms of orthogonality.

- Total Orthogonality [86] and [83]: The idea is to make one chroma channel zero and the other non-zero at any (u,v). For example, $\widetilde{h}_\alpha(u,v) = 0.9$ and $\widetilde{h}_\beta(u,v) = 0$.

- Quadrature Orthogonality [95]: The idea is to make one chroma channel real and the other imaginary at any (u,v), i.e., the two channels are modulated by a frequency $\boldsymbol{\omega}(u,v)$ but in quadrature phase. Translating the spatial domain, this means that

$$h_\alpha(m,n) = \gamma_c\sqrt{2}\cos\left(\boldsymbol{\omega}(u,v)^T \begin{bmatrix} m \\ n \end{bmatrix} - \phi\right) \tag{5.30a}$$

$$h_\beta(m,n) = \gamma_c\sqrt{2}\sin\left(\boldsymbol{\omega}(u,v)^T \begin{bmatrix} m \\ n \end{bmatrix} - \phi\right) \tag{5.30b}$$

  where $m \in \{0,\ldots,M-1\}$, $n \in \{0,\ldots,N-1\}$, and ϕ is the phase angle. In this way, the two channels can be easily separated during the demosaicking process using the orthogonality of cosine and sine functions.

We formulate the orthogonality criteria as a penalty function that is a surrogate of both approaches.

Definition 5.2.4 (Orthogonality Penalty) *For a CFA having chrominance channels with spectra $\widetilde{\boldsymbol{h}}_\alpha(\boldsymbol{\omega})$ and $\widetilde{\boldsymbol{h}}_\alpha(\boldsymbol{\omega})$, the orthogonality penalty is defined as*

$$\begin{aligned}\rho(\boldsymbol{h}_\alpha, \boldsymbol{h}_\beta) \stackrel{def}{=} &\sum_{u=0}^{M}\sum_{v=0}^{N}\Big(|\Re\widetilde{h}_\alpha(u,v)| + |\Re\widetilde{h}_\beta(u,v)|\Big) \\ &+ \sum_{u=0}^{M}\sum_{v=0}^{N}\Big(|\Im\widetilde{h}_\alpha(u,v)| + |\Im\widetilde{h}_\beta(u,v)|\Big)\end{aligned} \tag{5.31}$$

which can be written as a function in $\boldsymbol{x}$ as follows

$$\begin{aligned}\rho(\boldsymbol{x}) = &(\|\Re\boldsymbol{F}\boldsymbol{Z}_\alpha\boldsymbol{x}\|_1 + \|\Re\boldsymbol{F}\boldsymbol{Z}_\beta\boldsymbol{x}\|_1) \\ &+ (\|\Im\boldsymbol{F}\boldsymbol{Z}_\alpha\boldsymbol{x}\|_1 + \|\Im\boldsymbol{F}\boldsymbol{Z}_\beta\boldsymbol{x}\|_1)\end{aligned} \tag{5.32}$$

Looking at the first summation in (5.31), we notice that for every 2D frequency (u, v), the term $|\Re\widetilde{h}_\alpha(u,v)| + |\Re\widetilde{h}_\beta(u,v)|$ is the ℓ_1-norm of a 2-dimensional vector $[\Re\widetilde{h}_\alpha(u,v), \Re\widetilde{h}_\beta(u,v)]^T$. Therefore, minimizing this ℓ_1-norm promotes either one of the components to zero (or both). Similar argument applies for the imaginary components in the second summation. As a result, the total variation can be regarded as a proxy to the orthogonality condition.

5.3 Formulation of Optimal CFA Design Problem

Using the variables and constraints defined in the previous section, we present two different optimization formulations of the CFA design problem in this section: (i) A non-convex formulation that integrates all the above information into a single optimization, and (ii) convex relaxation that makes the problem more tractable.

5.3.1 Non-Convex CFA Design

The non-convex CFA optimization puts all the objectives and constraints defined in the previous section into a single optimization problem. This gives us

$$\begin{aligned}
&\underset{\boldsymbol{x}}{\text{maximize}} && \gamma_c(\boldsymbol{x}) + \lambda_l \gamma_l(\boldsymbol{x}) - \lambda_\rho \rho(\boldsymbol{x}) && && (5.33)\\
&\text{subject to} \\
& && \boldsymbol{x} \in [0,1]^{3L} && \text{(Realizability)} && \text{(a)}\\
& && (\boldsymbol{I} - \text{diag}(\boldsymbol{e}_1))\boldsymbol{F}\boldsymbol{Z}_l\,\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.2)} && \text{(b)}\\
& && \boldsymbol{U}\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.3)} && \text{(c)}\\
& && \boldsymbol{W}\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.4)} && \text{(d)}\\
& && \text{TV}(\boldsymbol{x}) \le \text{TV}_{\max} && \text{(Proposition 5.2.5)} && \text{(e)}
\end{aligned}$$

where λ_l and λ_ρ are the regularization parameters controlling the relative weights of the luminance sensitivity and the orthogonality penalty. The penalty function $\rho(\boldsymbol{x})$ is added to the objective with a negative sign so that it is minimized. By lower bounding $\gamma_c(\boldsymbol{x})$ with a constant τ, we can rewrite (5.33) as

$$\begin{aligned}
&\underset{\boldsymbol{x},\tau}{\text{maximize}} && \tau + \lambda_l \gamma_l(\boldsymbol{x}) - \lambda_\rho \rho(\boldsymbol{x}) && && (5.34)\\
&\text{subject to} \\
& && \boldsymbol{x} \in [0,1]^{3L} && \text{(Realizability)} && \text{(a)}\\
& && (\boldsymbol{I} - \text{diag}(\boldsymbol{e}_1))\boldsymbol{F}\boldsymbol{Z}_l\,\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.2)} && \text{(b)}\\
& && \boldsymbol{U}\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.3)} && \text{(c)}\\
& && \boldsymbol{W}\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.4)} && \text{(d)}\\
& && \text{TV}(\boldsymbol{x}) \le \text{TV}_{\max} && \text{(Proposition 5.2.5)} && \text{(e)}\\
& && \boldsymbol{x}^T\boldsymbol{Q}_\alpha\boldsymbol{x} \ge \tau^2 && \text{(Proposition 5.2.1)} && \text{(f)}\\
& && \boldsymbol{x}^T\boldsymbol{Q}_\beta\boldsymbol{x} \ge \tau^2 && \text{(Proposition 5.2.1)} && \text{(g)}
\end{aligned}$$

In this optimization problem, the objective and constraints are convex except for (5.34)(f) and (5.34)(g). This is because these inequalities include convex quadratic form in the "$\ge$" side, where convexity comes from the fact that $\boldsymbol{Q}_\alpha$ and $\boldsymbol{Q}_\beta$ are positive semidefinite matrices. Hence, the optimization problem is non-convex.

Algorithm 2 Successive Convex Approximations

Require: Initial guess $\boldsymbol{x}^{(0)}$, $k = 0$.

while γ_c not converge **do**

Replace the quadratic terms $\boldsymbol{x}^T\boldsymbol{Q}_\alpha\boldsymbol{x}$ and $\boldsymbol{x}^T\boldsymbol{Q}_\beta\boldsymbol{x}$ in inequalities (5.34)(f-g) by their first order Taylor approximations around $\boldsymbol{x}^{(k)}$:

$$\boldsymbol{x}^T\boldsymbol{Q}_\alpha\boldsymbol{x} \approx \boldsymbol{x}^{(k)T}\boldsymbol{Q}_\alpha\boldsymbol{x}^{(k)} + 2(\boldsymbol{x} - \boldsymbol{x}^{(k)})\boldsymbol{Q}_\alpha\boldsymbol{x}^{(k)} \geq \tau^2$$
$$\boldsymbol{x}^T\boldsymbol{Q}_\beta\boldsymbol{x} \approx \boldsymbol{x}^{(k)T}\boldsymbol{Q}_\beta\boldsymbol{x}^{(k)} + 2(\boldsymbol{x} - \boldsymbol{x}^{(k)})\boldsymbol{Q}_\beta\boldsymbol{x}^{(k)} \geq \tau^2$$

Solve the convex approximation of (5.34) to get $\gamma_c^{(k)}$

$k = k + 1$

end while

return $\boldsymbol{x}$

5.3.2 Solving the Optimization

While problem (5.34) is non-convex, we can find a local minimum by successive convex approximations [128]. The idea of successive convex approximation is to replace the quadratic terms in the non-convex constrains (5.34)(f) and (5.34)(g) by first order approximations around the initial guess $\boldsymbol{x}^{(0)}$. Since the quadratic form is convex, its first order approximation constitutes a lower bound. Hence, we are replacing the non-convex constraints (5.34)(f) and (5.34)(g) with convex but tighter constraints that limit the feasible set of $\boldsymbol{x}$. The algorithm repeats until τ converges to a fixed-point, which is the final chrominance sensitivity.

The overall algorithm is summarized in Algorithm 2. Figure 5.6 shows the convergence of Algorithm 2 for designing a 4×4 color atom. We notice the monotonic increase of τ until it converges to a fixed point. Since the original problem is non-convex, solution to the problem could be a local minimum depending on how the initialization is done. In practice, we solve the problem for multiple instances with different randomly generated initial guesses which approximately cover the design

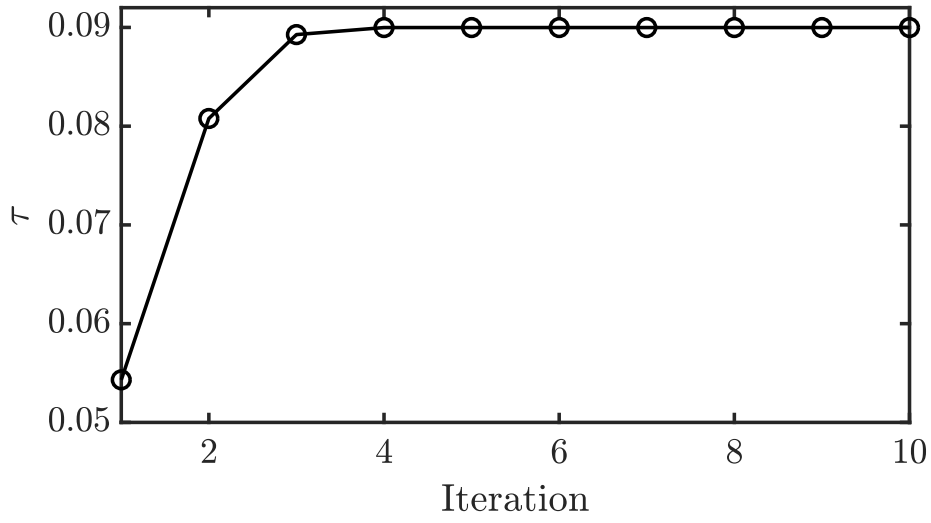


Fig. 5.6. Convergence of Algorithm 1 for 4×4 color filter design.

space (e.g., using the Latin hypercube sampling [129]), and pick the best solution among them.

5.3.3 Convex CFA Design

The relaxation from non-convex to convex can be done by explicitly forcing part of the chrominance components to zero. Specifically, we modulate the chrominance channels on the same frequency $\boldsymbol{\omega}(u,v) = (\frac{2\pi u}{M}, \frac{2\pi v}{N})$ using the quadrature orthogonality mentioned in (5.30). In terms of $\boldsymbol{x}$, these two equations can be written as:

$$\boldsymbol{Z}_\alpha \boldsymbol{x} = \gamma_c \boldsymbol{x}_c, \quad \boldsymbol{Z}_\beta \boldsymbol{x} = \gamma_c \boldsymbol{x}_s \tag{5.35}$$

where $\boldsymbol{x}_c$ and $\boldsymbol{x}_s$ are constant vectors that represent the vectorized version of the cosine and sine signals on the right hand side of (5.30a) and (5.30b), respectively, i.e.,

$$\boldsymbol{x}_c = \mathrm{vec}\left\{ \sqrt{2}\cos\left(\boldsymbol{\omega}(u,v)^T \begin{bmatrix} m \\ n \end{bmatrix} - \phi \right)_{m=0,n=0}^{M-1,N-1} \right\} \tag{5.36a}$$

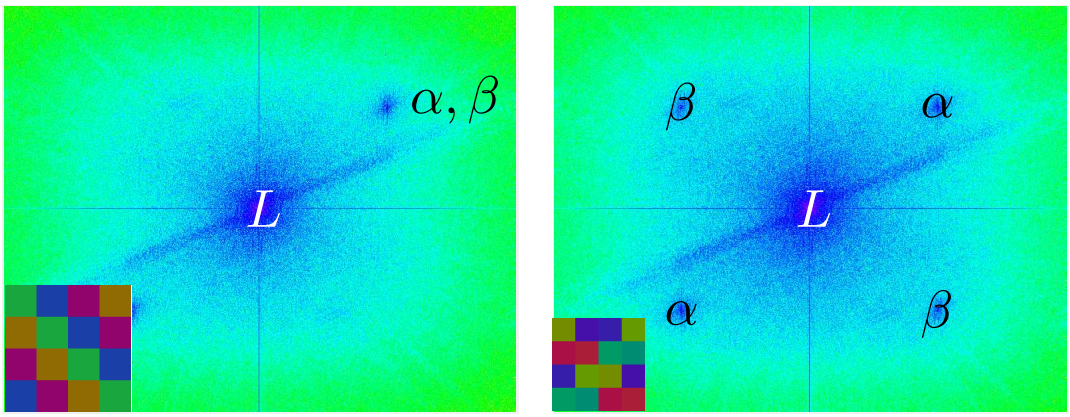
$$\boldsymbol{x}_s = \mathrm{vec}\left\{ \sqrt{2}\sin\left(\boldsymbol{\omega}(u,v)^T \begin{bmatrix} m \\ n \end{bmatrix} - \phi \right)_{m=0,n=0}^{M-1,N-1} \right\} \tag{5.36b}$$

Since we explicitly choose the modulation frequencies of chrominance channels manually, we can drop the aliasing constraint in Proposition 5.2.4. However, we still need the uniform luminance and chrominance constrains in Propositions 5.2.2 and 5.2.3. Moreover, since the luminance and chrominance gains are adversarial, the objective of this formulation is to maximize their weighted sum. To this end, the problem is written as:

$$\begin{aligned}
&\underset{\boldsymbol{x},\gamma_c}{\text{maximize}} && \gamma_c + \lambda_l \gamma_l(\boldsymbol{x}) && && (5.37)\\
&\text{subject to} && && &&\\
& && \boldsymbol{x} \in [0,1]^{3L} && \text{(Realizability)} && \text{(a)}\\
& && (\boldsymbol{I} - \text{diag}(\boldsymbol{e}_1))\boldsymbol{F}\boldsymbol{Z}_l\,\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.2)} && \text{(b)}\\
& && \boldsymbol{U}\boldsymbol{x} = \boldsymbol{0} && \text{(Proposition 5.2.3)} && \text{(c)}\\
& && \text{TV}(\boldsymbol{x}) \leq \text{TV}_{\max} && \text{(Proposition 5.2.5)} && \text{(d)}\\
& && \boldsymbol{Z}_\alpha \boldsymbol{x} - \gamma_c \boldsymbol{x}_c = 0 && && \text{(e)}\\
& && \boldsymbol{Z}_\beta \boldsymbol{x} - \gamma_c \boldsymbol{x}_s = 0 && && \text{(f)}
\end{aligned}$$

In our terminology, the optimization problem of [95] is obtained from (5.37) by removing the crosstalk constraint (5.37)(d). Hence, our optimization limits the search space of the optimization in [95] to get CFAs that have acceptable crosstalk performance.

Figure 5.7 shows two color atoms obtained using the convex and non-convex formalizations. In the convex formulation, we select the modulation frequency to be $\omega_0 = [\pi, \pi]$ and the phase that maximizes γ_c at this frequency is found to be $\phi = \pi/12$. Then, we solve the problem to get $(\gamma_l, \gamma_c, TV) = (0.573, 0.08, 0.263)$. As for the non-convex formulation, we let the optimization to choose modulation frequencies subject to crosstalk and aliasing constraints. Solving the non-convex formulation yields $(\gamma_l, \gamma_c, TV) = (0.573, 0.09, 0.263)$. We notice that the non-convex formulation achieves higher chrominance sensitivity because of its flexibility in choosing the modulation frequencies.



(a) Convex Formulation

$\gamma_c(\boldsymbol{x}) = 0.08$

(b) Non-Convex Formulation

$\gamma_c(\boldsymbol{x}) = 0.09$

Fig. 5.7. 4×4 color atoms and corresponding spectra obtained using convex and non-convex formulations. Spectra are obtained from mosaicking the "Bikes" image in Kodak color dataset by the color atoms. Both have the same luminance sensitivity $\gamma_l(\boldsymbol{x}) = 0.577$ and same Total variation $TV(\boldsymbol{x}) = 0.263$.

5.4 Universal Demosaicking

In this section, we present a universal demosaicking algorithm which can be used to all CFAs presented in this paper. Our algorithm comprises two main parts: (i) a demosaicking step to remove the color filtering effect (Section 5.4.2) and (ii) a color correction step to mitigate the crosstalk effect (Section 5.4.3).

5.4.1 Special Consideration for QIS.

Before we talk about the demosaicking algorithm, we should briefly discuss the photon statistics of QIS. In CMOS, the measured voltage can be modeled as a nominate value corrupted by i.i.d. Gaussian noise. In QIS, previous work showed that the measured photon counts follow a truncated Poisson process [24]. When averaging over a number of temporal frames, the truncated Poisson becomes a Binomial [69]. If the photon count is sufficiently high, this binomial will approximately approaching

to a Gaussian. Applying the law of large numbers on the distribution of $\boldsymbol{B}$ in (5.5), the average is

$$\frac{1}{T}\sum_{t=0}^{T-1} b_{m,t} \xrightarrow{a.e.} \mathbb{E}[B_m] = 1 - \Psi_q(\theta_m),$$

and the maximum-likelihood estimate of the signal is

$$\theta_m = \Psi_q^{-1}\left(1 - \frac{1}{T}\sum_{t=1}^{T} b_{m,t}\right)$$

As discussed in [69], we can regard this equation as a tone-mapping of the photon counts. We regard θ_m as the m-th pixel of the mosaicked image generated by the CFA. The goal of demosaicing is to reconstruct a color image from θ_m.

5.4.2 Demosaicking by Frequency Selection

Our demosaicking algorithm is based on frequency selection [73]. It generalizes [74] as it works for any CFA as long as it satisfies the orthogonality constraints in Section 5.2.5

The key idea of the algorithm is to shift every chrominance channel to the base-band by multiplying with its carrier, then use a low-pass filter to reconstruct it. For chrominance components that are replicated over distinct carriers, we combine them by simple averaging. After obtaining the α and β chrominance channels, they are re-modulated to their original positions and subtracted from the mosaicked image to obtain the luminance channel. This process is summarized in Algorithm 3 for a special case of a CFA that has strictly one replica of the α and β chrominance channels. It is also illustrated by Figure 5.8. Extension of the algorithm to the general case is straightforward.

To apply Algorithm 3 on CFAs proposed in [86], [93] and [98], they must satisfy the orthogonality constraints in Section 5.2.2. However, this is not satisfied with our choice of the luminance/chrominance basis defined by $\boldsymbol{T}$ in (5.7). Hence, we use for every CFA the transformation matrix $\boldsymbol{T}$ that makes its luminance and chrominance channel orthogonal. To ensure fairness, we normalize the matrix rows to unity so

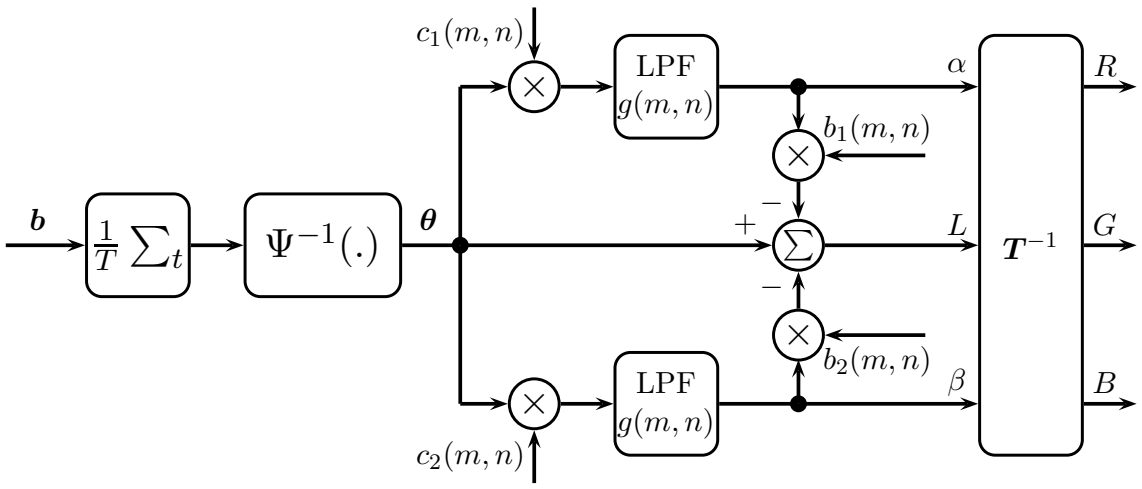


Fig. 5.8. Illustration of Algorithm 3 of demosaicking by frequency selection for a special case of a CFA that has strictly one replica of the α and β chrominance channels. Variable on the figure are defined in Algorithm 3.

that all luminance and chrominance have the same noise power. The transformation matrices are provided in the supplementary.

5.4.3 Color Correction

The demosaicking algorithm in Algorithm 3 does not take into account of the crosstalk effect. Like most of the mainstream image and signal processing (ISP) pipelines, we reduce the cross-talk via a color correction step.

Given the demosaicked color pixel $\widehat{\mathbf{im}}(m,n)$, the color correction multiplies $\widehat{\mathbf{im}}(m,n)$ by a 3×3 matrix $\boldsymbol{M}$ such that $\boldsymbol{M}\widehat{\mathbf{im}}(m,n)$ is the color-corrected pixel. The matrix $\boldsymbol{M}$ is learned by comparing a measured color pixel to a known color chart value. Mathematically, suppose we have K measured color pixels forming a $3\times K$ matrix $\boldsymbol{Q}_{\text{False}}$, and a corresponding true color values forming another $3\times K$ matrix $\boldsymbol{Q}_{\text{GT}}$, we can estimate $\boldsymbol{M}$ by solving

$$\boldsymbol{M} = \arg\min_{\boldsymbol{M}} \epsilon_c(\boldsymbol{M}) \text{ subject to } \boldsymbol{M}\boldsymbol{u} = \boldsymbol{u} \tag{5.38}$$

Algorithm 3 Demosaicking by Frequency Selection

Require: The image $\boldsymbol{\theta}$ which is mosaicked by a CFA of size $M \times N$ as defined in (5.8), a luminance/chrominance transformation matrix $\boldsymbol{T}$, a low-pass filter g, a scalar $K \stackrel{\text{def}}{=} MN$ and a scalar $r = \begin{cases} 2 & \text{if } \boldsymbol{\omega} = (\pi, \pi) \\ 1 & \text{otherwise} \end{cases}$.

Ensure: α and β chrominance channels are modulated on carriers $\boldsymbol{\omega}(u_1, v_1)$ and $\boldsymbol{\omega}(u_2, v_2)$, respectively.

1) Reconstruct the α chrominance channel

$$\alpha(m,n) = (\theta(m,n)c_1(m,n)) \circledast g(m,n)$$

where

$$c_1(m,n) = \frac{K}{|a_1|} \cos\left(\boldsymbol{\omega}(u_1,v_1)^T \begin{bmatrix} m \\ n \end{bmatrix} + \angle a_1 \right)$$

and $a_1 = \widetilde{h}_\alpha(u_1, v_1)$

2) Reconstruct the β chrominance channel

$$\beta(m,n) = (\theta(m,n)c_2(m,n)) \circledast g(m,n)$$

where

$$c_2(m,n) = \frac{K}{|a_2|} \cos\left(\boldsymbol{\omega}(u_2,v_2)^T \begin{bmatrix} m \\ n \end{bmatrix} + \angle a_2 \right)$$

and $a_2 = \widetilde{h}_\beta(u_2, v_2)$

3) Reconstruct the luminance channel

$$L(m,n) = \theta(m,n) - \alpha(m,n)b_1(m,n) - \beta(m,n)b_2(m,n)$$

where

$$b_1(m,n) = \frac{2|a_1|^2}{rK^2} c_1(m,n) \text{ and } b_2(m,n) = \frac{2|a_2|^2}{rK^2} c_2(m,n)$$

return $[\boldsymbol{R}, \boldsymbol{G}, \boldsymbol{B}]^T = \boldsymbol{T}^{-1}[\boldsymbol{L}, \boldsymbol{\alpha}, \boldsymbol{\beta}]^T$

Before Color Correction After Color Correction

Fig. 5.9. Effect of color correction on retaining vivid image colors.

where $\epsilon_c(\boldsymbol{M}) = \mathrm{Tr}\left\{(\boldsymbol{M}\boldsymbol{Q}_{\mathrm{False}} - \boldsymbol{Q}_{\mathrm{GT}})^T (\boldsymbol{M}\boldsymbol{Q}_{\mathrm{False}} - \boldsymbol{Q}_{\mathrm{GT}})\right\}$ is the color error. $\boldsymbol{u} \overset{\mathrm{def}}{=} [0.95, 1, 1.0889]^T$ is the white point for D65 illuminant. To minimize the noise amplification, it is advised to add regularization when estimating $\boldsymbol{M}$ [130]. Since a standard color chart comprises 24 color patches, we can estimate the noise by computing the norm of covariance matrix of every color patch, and get the average value over the 24 color patches. Hence, the optimization problem is rewritten as

$$\begin{aligned} \boldsymbol{M} &= \arg\min_{\boldsymbol{M}} \epsilon_c(\boldsymbol{M}) + \kappa \sum_{i=1}^{24} ||\mathrm{Cov}(\boldsymbol{M}\boldsymbol{Q}_{\mathrm{False}}^{(i)})||_2^2 \\ &\text{subject to } \boldsymbol{M}\boldsymbol{u} = \boldsymbol{u} \end{aligned} \tag{5.39}$$

where $\boldsymbol{Q}_{\mathrm{False}}^{(i)}$ represents the pixels of the ith color patch, and κ is a positive scalar that controls the noise amplification effect. By varying κ, we can draw a tradeoff curve between color reproduction accuracy and noise amplification.

Figure 5.9 shows reconstructed images before and after color correction. In this figure, the crosstalk parameters are $(\alpha_r, \alpha_r, \alpha_r) = (0.23, 0.15, 0.1)$. We can see the effect of color correction in the more saturated red and yellow feathers and in the green leaves in the background.

6. EXPERIMENTAL EVALUATION

In this chapter, we present our experimental results for the QIS image reconstruction and threshold design problems. On the image reconstruction side, we study the convergence of ADMM algorithm in Section 6.1, then we present in Section 6.2 a comparison between our proposed Transform-Denoise algorithm and other algorithms. For performance evaluation, we use the peak signal-to-noise ratio (PSNR) for assessing the reconstruction quality, and we use the elapsed CPU time as a proxy for assessing the computational complexity. On the threshold design side, we evaluate our proposed threshold update scheme by comparing it with existing methods. We consider two evaluation metrics: (1) convergence rate of the threshold update methods; (2) quality of the reconstructed images. For regular imaging experiments, we use our own Purdue dataset comprising 77 images captured by a Canon EOS Rebel T6i camera. For HDR imaging experiments, we use the HDR-Eye dataset by Nemoto et al. [131, 132].

6.1 Convergence of ADMM Reconstruction Algorithm

In this experiment, we test the convergence of the ADMM algorithm used to get the ML solution (subsection 3.1.1) and the MAP solution (Section 3.2). We choose the MAP solution with TV prior because we can calculate the objective function since the prior term is explicitly defined ($||\boldsymbol{Dc}||_1$). We could not do that with the Plug-and-Play algorithm because the prior term is implicitly defined by the denoiser $\mathcal{D}$. QIS simulation parameters are $q = 1$, $K = 4$, $T = 5$, and $\alpha = 2K^2$.

We assume the interpolation filter $\{g_k\}$ has a box-car kernel. As a result, we can use the ML closed-form in Proposition 3.1.2 which gives the exact unique ML solution that the ADMM algorithm must converge to (because the problem is convex). Table 6.1 shows the reconstruction PSNR and CPU time of the ML solution obtained

by applying the ADMM algorithm with $\rho = 10$ (ML-ADMM) and that obtained by the closed-form expression (ML-CF). Compared to ML-ADMM that can compute an approximate ML estimate in 196 seconds using 40 ADMM iterations [1], the closed-form expression can compute the exact ML estimate in a fraction of second.

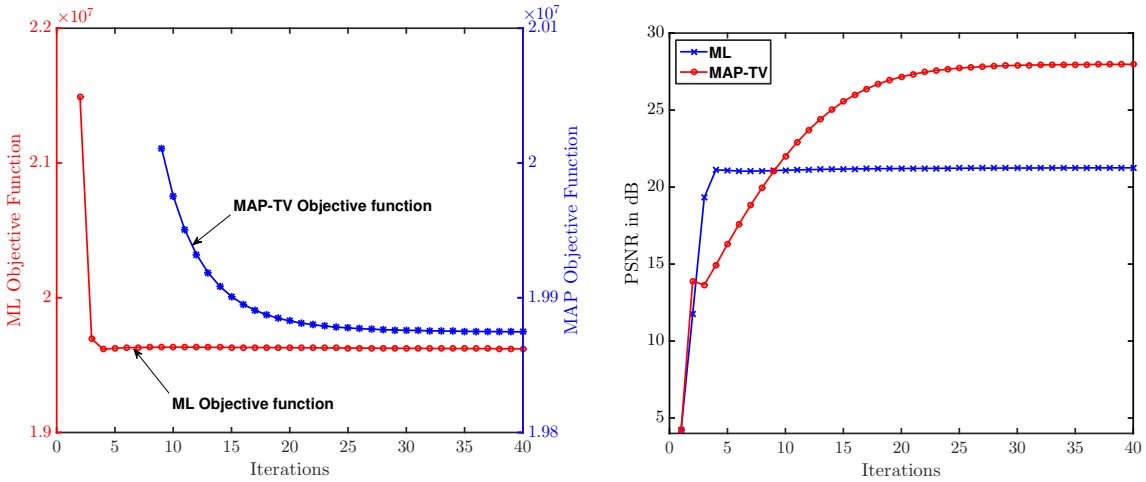
Table 6.1.
Reconstruction PSNR in dB and CPU time in seconds for ML solution. Both values are averaged on 77 images in our dataset.

Algorithm	ML-ADMM	ML-CF
PSNR (dB)	21.99	22.02
CPU Time (Sec)	196.24	0.46

Figure 6.1 shows the convergence of the ML-ADMM algorithm and the MAP-TV-ADMM algorithm with for TV prior. For both algorithms $\rho = 10$, and for the MAP-TV-ADMM algorithms $(\lambda, \gamma) = (5, 35)$ which are obtained by exhaustive search on a grid of possible values. The optimization criterion of this exhaustive search is the PSNR after 40 iteration. We notice that both algorithms converge to a unique solution, where the MAP-ADMM algorithm has a slower convergence rate. In addition, the reconstruction PSNR increases slowly with iterations in the MAP-TV-ADMM case to reach a value higher than the ML-ADMM solution by 6.7 dB.

Figure 6.2 shows the reconstructed images using the ML Closed-Form expression, the ML-ADMM estimate, and the MAP-TV-ADMM estimate. We notice that the ML-CF and ML-ADMM images are nearly the same, and the MAP-TV-ADMM image is better than both of them.

[1] Theoretically, the exact ML estimate is obtained if we run the ADMM algorithms for large number of iterations until the likelihood function converges with a sufficiently high numerical precision.



(a) Convergence (b) PSNR

Fig. 6.1. Simulated QIS data and the reconstructed gray-scale images using different reconstruction methods.

(a) 21.27 dB (b) 21.25 dB (c) 27.97 dB

Fig. 6.2. Reconstructed Images using ML closed-form (a) and ML ADMM algorithm (b) are nearly the same. The image reconstituted using MAP-TV ADMM algorithm (c) is better than both.

6.2 Image Reconstruction Performance

In this experiment, we compare between our Transform-Denoise (TD) algorithm and other algorithms. For performance evaluation, we compute the reconstruction PSNR and CPU time averaged on 77 images in our dataset. QIS simulation parameters are $q = 1$, $K = 4$, $T = 5$, and $\alpha = 2K^2$

As mentioned before, the pure ML solution is not useful because it is too noisy, and a denoising step is necessary as the TD algorithm suggests. In this experiment,

we try two different denoisers in our TD algorithm: 1) a learning-based denosier [133] which is based on the training of a deep convolutional neural network (CNN), and 2) the BM3D denoiser [105].

We compare the TD algorithms with three different MAP solutions obtained by different priors: 1) MAP solution with total-variation prior [22] (MAP-TV), 2) MAP solution with BM3D denoiser prior (MAP-TV), and 3) MAP solution with CNN denoiser prior. The MAP-TV solution is obtained by applying 40 iterations of the ADMM algorithm; whereas the MAP-BM3D and MAP-CNN solutions are obtained by applying 40 iterations of the Plug-and-Play algorithm. For all MAP solutions, the value of $\rho = 10$, and the values of γ and λ are obtained by exhaustive search on a grid of possible values. The optimization criterion of this exhaustive search is the PSNR after 40 iteration. The optimized parameters are $(\lambda, \gamma) = (5, 35)$ for MAP-TV, $(\lambda, \gamma) = (2, 70)$ for MAP-BM3D, and $(\lambda, \gamma) = (5, 60)$ for MAP-CNN.

Table 6.2.
Reconstruction PSNR in dB and CPU time in seconds for MAP solution and the TD solutions. Both values are averaged on 77 images in our dataset.

Algorithm	MAP-TV [22]	MAP-BM3D	TD-BM3D	MAP-CNN	TD-CNN
PSNR (dB)	28.55	29.71	30.43	30.04	30.29
CPU Time (Sec)	197.47	524.67	6.87	267.71	2.33

As shown in Table 6.2, the TD algorithm achieves the best reconstruction quality in terms of PSNR in much less time than the iterative ADMM and Plug-and-Play algorithms. This is intuitive because the TD algorithm in non-iterative and other algorithms are iterative. We emphasize that ADMM and Plug-and-Play algorithms can obtain better PSNR than these values if we run them for more than 40 iterations or if we fine tune the parameters λ, γ, and ρ by exhaustive search on a fine grid of suggested parameters. On the other hand, the TD algorithm is parameter-free because the noise level after Anscombe transformation is fixed and known.

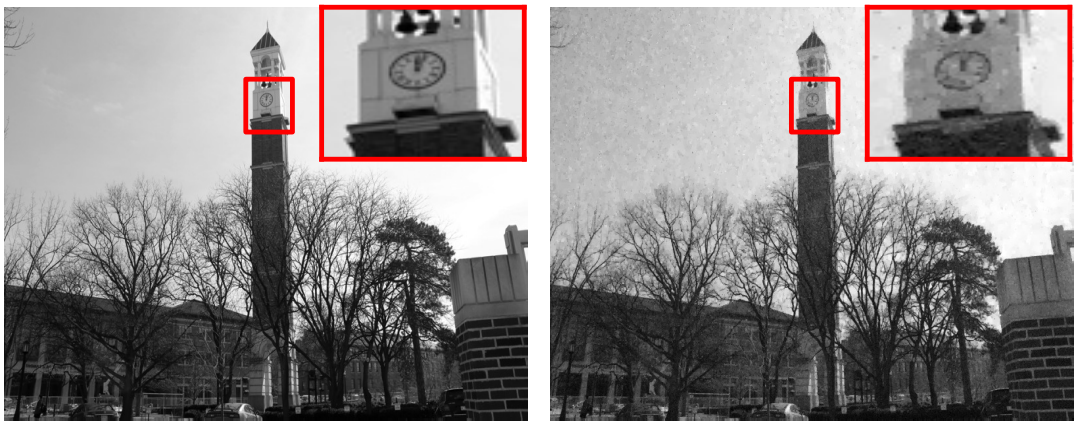
Figure 6.3 and Figure 6.4 show reconstructed image by different algorithms compared to the ground truth. We notice that the TD algorithm achieves the highest visual quality compared to other algorithms. It is worth noting that the TD-CNN algorithm can reconstruct more details than the TD-BM3D algorithm. This is attributed to the high learning capacity of CNNs which is trained on dataset comprising thousands of images. This leads to more powerful prior term compared to BM3D.

6.3 Convergence of The Threshold Update Scheme

We compare the proposed threshold update scheme with the Markov Chain (MC) adaptation proposed by Hu and Lu [62]. The Markov Chain adaptation models the threshold as a variable with 2^L states. These 2^L states can be regarded as 2^L steps before reaching to the next threshold level. The probability of changing from one state to another is controlled by a parameter $1-\beta$ with $0 < \beta < 1$. When a bit arrives, the state will be updated (increased or decreased) or will remain unchanged. Once the state is increased by 2^L times, the threshold will be increased by one.

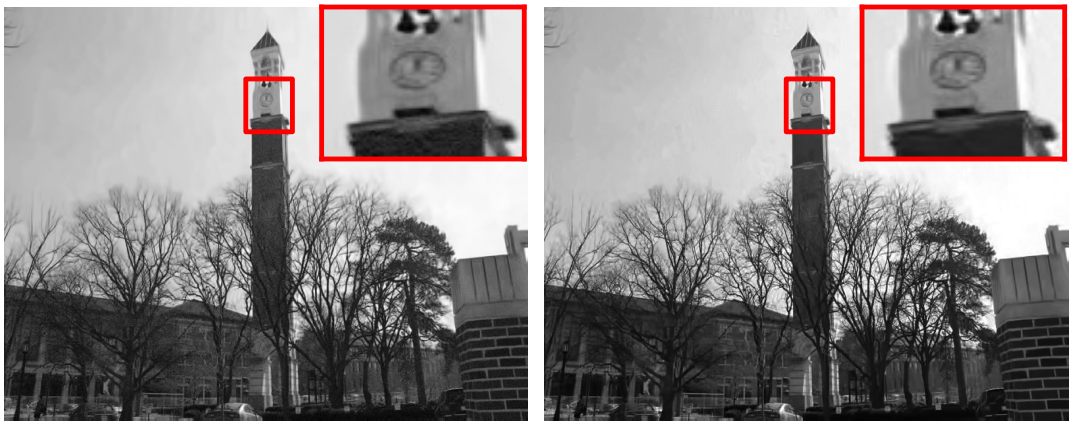
When comparing Markov Chain adaptation with the proposed bisection algorithm, one should be aware of the difference between the two methods. Markov Chain adaptation is a *per-jot* update scheme whereas the proposed bisection algorithm is a *per-pixel* update scheme. For a pixel with $K \times K$ jots, Markov Chain adaptation needs K^2 iterations to update the threshold *sequentially*. In contrast, the proposed bisection algorithm updates a common threshold for all K^2 jots *simultaneously*. Thus in practice our bisection algorithm is significantly less complex to implement in hardware than the Markov Chain. In order to take the different forms of updates into account, we treat the K^2 iterations of Markov Chain adaptation as one "major iteration" and compare it with the one bisection step of the proposed algorithm.

The first comparison we make is to check the threshold at different jots. Figure 6.5 shows the results of three typical runs with underlying optimal thresholds $q^* = 1, 8, 16$.



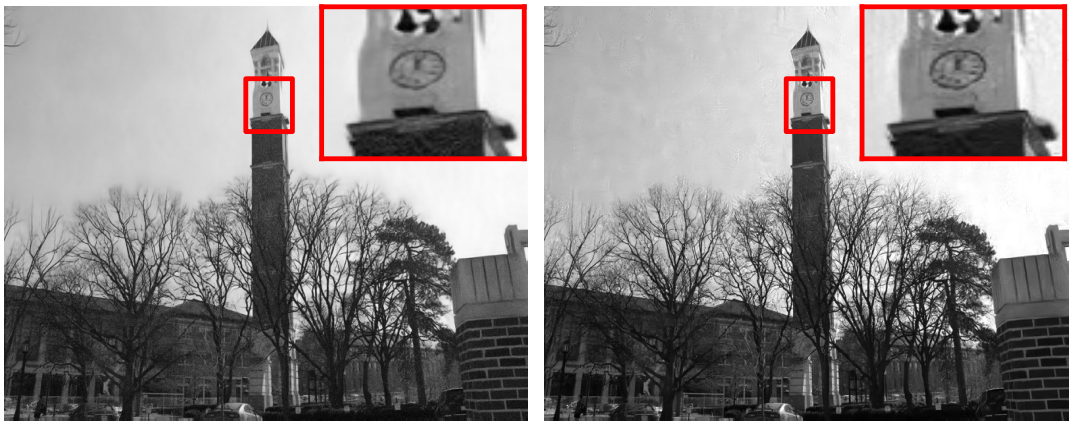
(a) Ground Truth

(b) MAP-TV [22], 40 iter., 28.22 dB, 194 sec



(c) MAP-BM3D, 28.62 dB, 515 sec

(d) TD-BM3D, 29.72 dB, 7 sec



(e) MAP-CNN, 29.20 dB, 265 sec

(f) TD-CNN, 29.77 dB, 2.4 sec

Fig. 6.3. Simulated QIS data and the reconstructed gray-scale images using different reconstruction methods.

(a) Ground Truth

(b) MAP-TV [22], 40 iter., 21.99 dB, 195 sec

(c) MAP-BM3D, 27.44 dB, 515 sec

(d) TD-BM3D, 28.17 dB, 6 sec

(e) MAP-CNN, 28.01 dB, 269 sec

(f) TD-CNN, 28.53 dB, 2 sec

Fig. 6.4. Simulated QIS data and the reconstructed gray-scale images using different reconstruction methods.

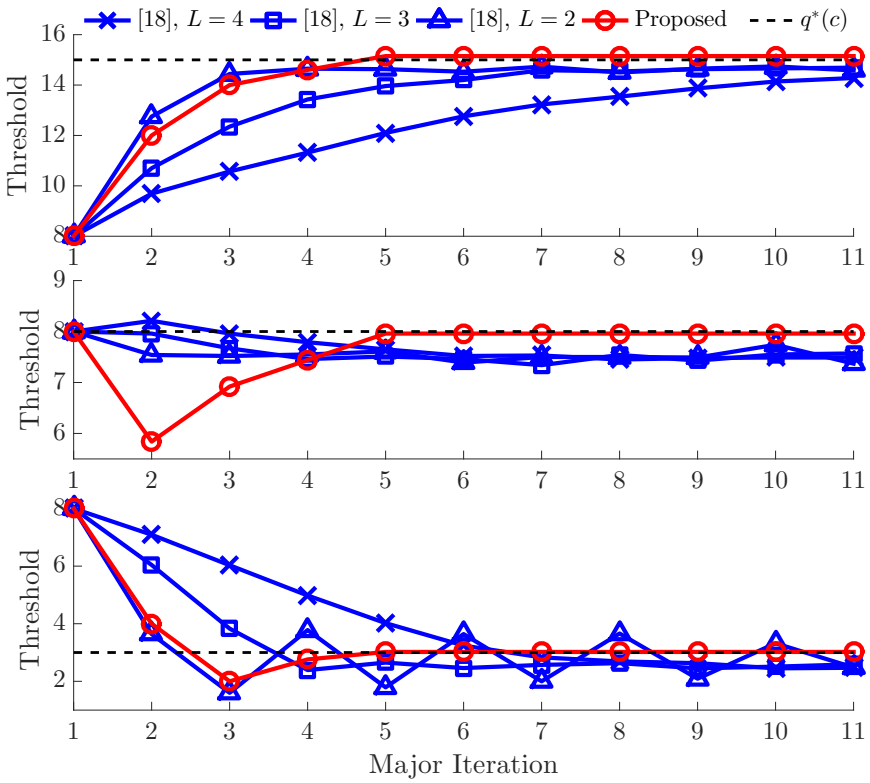


Fig. 6.5. Convergence of the threshold at 3 jots. Each curve is averaged over 100 random samples. The red curve indicates the proposed bisection method. The black curves are the Markov chain adaptation [62] with $\beta = 0.25$. Note that one major iteration of Markov Chain adaptation corresponds to K^2 sequential updates, and one major iteration of the bisection method corresponds to a single update to K^2 jots simultaneously.

In this experiment, we generate 100 random binary blocks of size $K \times K$ and estimate the threshold at each major iteration. We report the average of these 100 estimates to minimize the randomness of the data. The results show that one iteration of the proposed bisection algorithm works as good as the K^2 iterations of the Markov Chain adaptation. In some cases, Markov Chain tends to oscillate whereas the bisection result is stable.

The second comparison we make is to check how close the estimated threshold is compared to the optimal threshold. The optimal threshold q^* is obtained using the oracle scheme. In Figure 6.6, we plot the mean squared error between the estimated

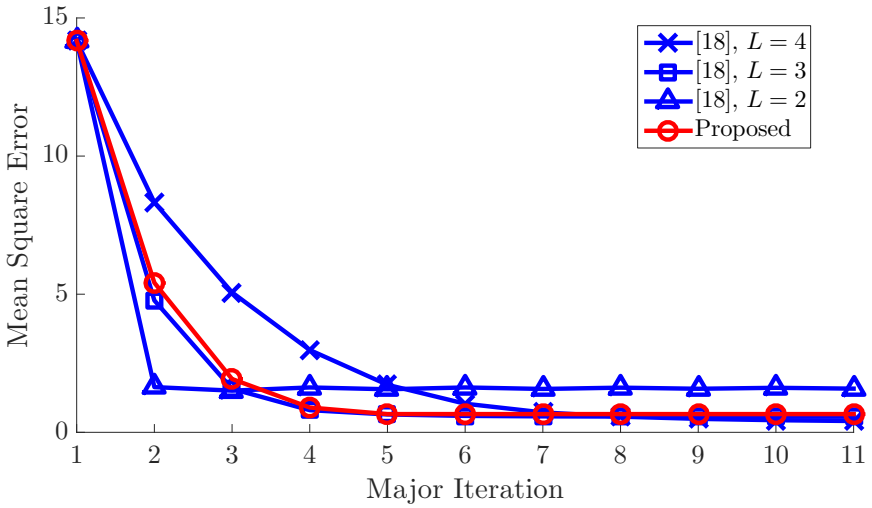


Fig. 6.6. Mean square error between the estimated threshold and the ideal oracle threshold. Each curve is averaged over 50 random samples and 77 images. The red curve indicates the proposed bisection method. The black curves are the Markov chain adaptation [62] with $\beta = 0.25$.

threshold and the oracle threshold. For fairness, we show the results of the MSE averaged over the 77 images of our dataset, and 50 random samples per image. One threshold is shared by $K \times K$ jots, and each $K \times K$ jots correspond to one pixel. The result is consistent with the ones shown in Figure 6.5.

6.4 Influence of QIS Threshold on Image Reconstruction Quality

The convergence comparison in the previous subsection is only useful to compare threshold update methods that actually return a threshold. In the QIS literature, there are methods that implicitly update the threshold, e.g., the conditional reset method [45]. For comparison with these methods, we have to compare the quality of the image reconstructed from the binary raw data. The image reconstruction is done using the closed-form ML estimate in Proposition 3.1.2.

In this experiment, we fix the spatial over-sampling factor as $K = 4$, and number of temporal frames as $T = 13$. The maximum threshold level is set as $q_{\max} = 16$ to ensure that it is realistic for today's QIS, and $\alpha = 15K^2$.

We consider three classes of methods:

- Uniform Threshold. Uniform threshold is commonly used in the device literature [20–22]. A uniform threshold is a single threshold applied to all pixels in the image. In this experiment, we consider the following choices of uniform thresholds: $q = 1$, $q = 5$, $q = 10$ and $q = 16$.

- Conditional Reset [45]. Conditional reset counts the number of photons and is reset when it is above the threshold. The threshold in conditional reset is sequentially increasing or decreasing. The reconstructed image is obtained by digitally integrating the raw binary frames.

- Proposed Method. As we discussed in Section 4.2.5, the proposed method can be implemented to let multiple pixels share a common threshold. Thus, in this experiment we consider three sharing strategies: (1) Share a threshold between a neighborhood of $K \times K$ jots (i.e., one threshold for one pixel); (2) Share a threshold between a neighborhood of $K^2 \times K^2$ jots (i.e., one threshold for $K \times K$ pixel); (3) Share a threshold between a neighborhood of $2K^2 \times 2K^2$ jots (i.e., one threshold for $2K \times 2K$ pixels).

The result of the experiment is shown in Table 6.3. The PSNR values reported are averaged over 77 images in our dataset. Each image generates 50 random realizations, and the PSNR of an image is averaged over these 50 random realizations to minimize the randomness. As shown in the table, while conditional reset generally performs better than a uniform threshold, it performs significantly worse than the proposed threshold update scheme.

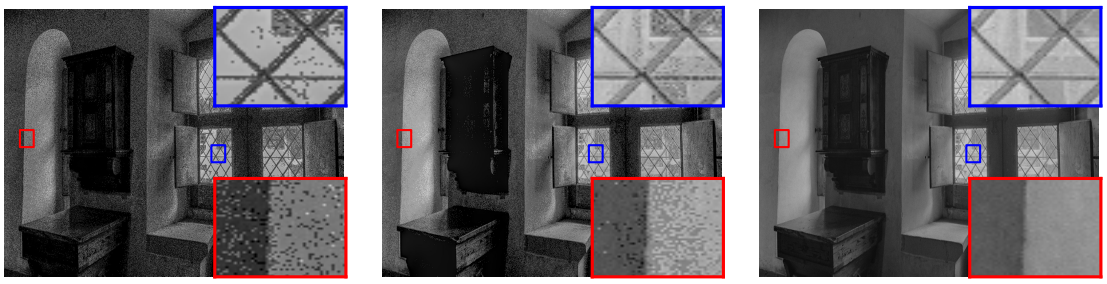
Table 6.3.
Average PSNR and Standard deviation of 77 recovered images using different Q-maps and 50 random samples.

	Configuration	Average PSNR	Std
Uniform Threshold	$q = 1$	10.30	0.01
	$q = 5$	28.80	0.04
	$q = 10$	23.22	0.02
	$q = 16$	12.95	0.01
Conditional Reset [45]	Ascending q sequence	23.77	0.52
	Descending q sequence	24.95	0.53
Proposed Method	$2K^2 \times 2K^2$	30.14	0.06
	$K^2 \times K^2$	31.18	0.06
	$K \times K$	32.78	0.02

Fig. 6.7. Bracketed images with different exposure settings. From Left to Right: -2.7,-2, -1.3, -0.7, 0, 0.7, 1.3, 2, and 2.7 EV.

6.5 Influence of QIS Threshold on HDR Imaging

Since QIS does not have sufficient full well capacity to accumulate photons for HDR imaging, we apply the dynamic range extension method discussed in Section 4.2.6. When different threshold schemes are used, the reconstructed HDR images will be affected. The objective of this experiment is to evaluate the influence of the threshold in HDR imaging.



$q = 1$, PSNR = 17.94 dB $q = 16$, PSNR = 20.77 dB Proposed, PSNR = 31.46 dB

Fig. 6.8. The reconstructed HDR images using different thresholds. See supplementary material for additional results.

In this experiment, we consider the HDR-Eye image dataset [131, 132]. Each HDR image in this dataset contains 9 images acquired at different exposure settings (-2.7,-2, -1.3, -0.7, 0, 0.7, 1.3, 2, and 2.7 EV). A snapshot of these images are shown in Figure 6.7. From each exposure, we simulate the photon counts resulting from the luminance channel. The sensor gain is set as $\alpha = K^2(q_{\max} - 1)$ to ensure proper number of photons, where $K = 4 \times 4 = 16$ and $q_{\max} = 16$. On the reconstruction side, we reconstruct the 9 images using the ML solution presented in Proposition 3.1.2. Tone mapping and exposure fusion [12] are applied to the 9 images to generate an HDR image. As a reference, we apply the same tone mapping and fusion algorithm to the 9 ground truth images. PSNR between the reference and the estimated is then recorded. QIS simulation parameters are $K = 4$, $T = 13$, $q_{\max} = 16$, and $\alpha = 15K^2$.

The result of this experiment is shown in Figure 6.8. With the proposed threshold update scheme, the reconstructed images achieve the highest PSNR value and visual quality. When $q = 1$, which is too low, the image appears under-exposed. When $q = 16$, which is too high, the image appears over-exposed. The spatially varying property of the proposed method mitigates the issue by allowing multiple thresholds.

In practice, one would typically add image denoisers to handle the randomness in the ML estimate and potentially other types of noise. This can be done using methods

such as [24]. In HDR literature, there are also optical approaches that reduce the number of exposures, e.g., [134, 135]. These techniques are complementary to QIS, because QIS is a sensor of similar functionality of a CMOS. Thus optical techniques can always be added.

In this section, we present CFAs obtained using our optimization framework in Section 6.6. Afterwards, we evaluate the performance of different CFAs using the universal demosaicking algorithm proposed in Section 5.4. First, using the Macbeth color chart, we show the noise-color trade-off of our robust CFAs comparsed to others in Section B.4. Second, we show in Section B.3 a quantitative and qualitative comparison of the reconstruction performance of all CFAs on Kodak [136] and McMaster [137] color datasets.

6.6 Proposed Solutions of CFA Design Problem

We focus on the non-convex formulation (5.34) since it is more flexible than the convex formulation (5.37). We set the parameters of (5.34) as $\lambda_c = 10$ and $\lambda_\rho = 0.02$. We run multiple instances of Algorithm 2 (2000 instances) using different random initializations for the color atoms $\boldsymbol{x}^{(0)}$. Then, we pick the solution with the highest chrominance sensitivity. To ensure that the initial guess spans the feasible set of $\boldsymbol{x}$, we use uniform Latin hypercube sampling of the domain $[0, 1]^{3L}$.

Figure 6.9 shows our proposed CFAs and their accompanied spectra compared to other CFAs in the literature. For every CFA, we compute 1) the luminance and chrominance gains (γ_l and γ_c) in (5.17) to measure robustness to noise, 2) the total variation metric $TV(x)$ (Proposition 5.2.5) to measure robustness to crosstalk, and 3) the aliasing metric J_l in (5.23) to measure robustness to aliasing, and the condition number $\kappa(\boldsymbol{A})$ defined in Section 5.2.4. To calculate the aliasing metric for [83], [86], [93] and [98], we use the transformation matrices that make the luminance and chrominance channels orthogonal as mentioned in Section 5.4.2. Results are summarized in Table B.1.

Table 6.4.
CFA parameters and Reconstruction quality measured by YSNR and SMI metrics on Macbeth ColorChecker and average CPSNR on Kodak and McM color datasets. An arrow is placed after each metric to show whether it should be increased or decreased

Size	CFA Pattern	CFA Parameters					YPSNR ↑		SMI ↑		CPSNR ↑	
		$\gamma_l \uparrow$	$\gamma_c \uparrow$	TV ↓	$J_l \downarrow$	$\kappa(\boldsymbol{A}) \downarrow$	w/o Ctk	w/ Ctk	w/o Ctk	w/ Ctk	w/o Ctk	w/ Ctk
4×4	RGBCY [93]	0.679	**0.125**	0.313	358	2.567	23.63	27.84	92.38	91.61	28.85	28.74
	RGBCWY [93]	0.597	0.102	**0.263**	329	2.328	23.43	26.34	91.18	83.49	28.58	28.41
	Hao *et* al. [86]	0.586	0.094	0.413	**93**	2.791	**24.45**	23.13	93.82	92.05	**30.76**	30.12
	Ours	0.577	0.090	**0.263**	147	**2.264**	24.35	**28.05**	**93.94**	**95.04**	30.32	**30.24**
3×3	Cheng *et* al. [98]	0.612	**0.167**	0.350	264	**2.132**	**24.50**	**27.77**	92.70	91.46	29.72	29.67
	Ours	0.577	0.115	0.350	**97**	2.201	24.48	26.80	**94.02**	**94.93**	**30.61**	**30.44**
3×2	Condat [95]	0.866	**0.250**	0.633	181	**2.151**	**25.29**	**26.40**	94.34	94.51	30.76	30.47
	Ours	0.866	0.187	**0.554**	**157**	2.484	25.20	26.19	94.34	**94.55**	**30.87**	**30.51**
4×2	Hirakawa-Wolfe [83]	0.866	0.125	0.550	**173**	3.197	**25.43**	25.91	94.22	93.88	**30.85**	30.33
	Ours	0.866	**0.187**	**0.513**	181	**2.369**	25.08	**27.11**	**94.30**	**94.99**	30.84	**30.62**

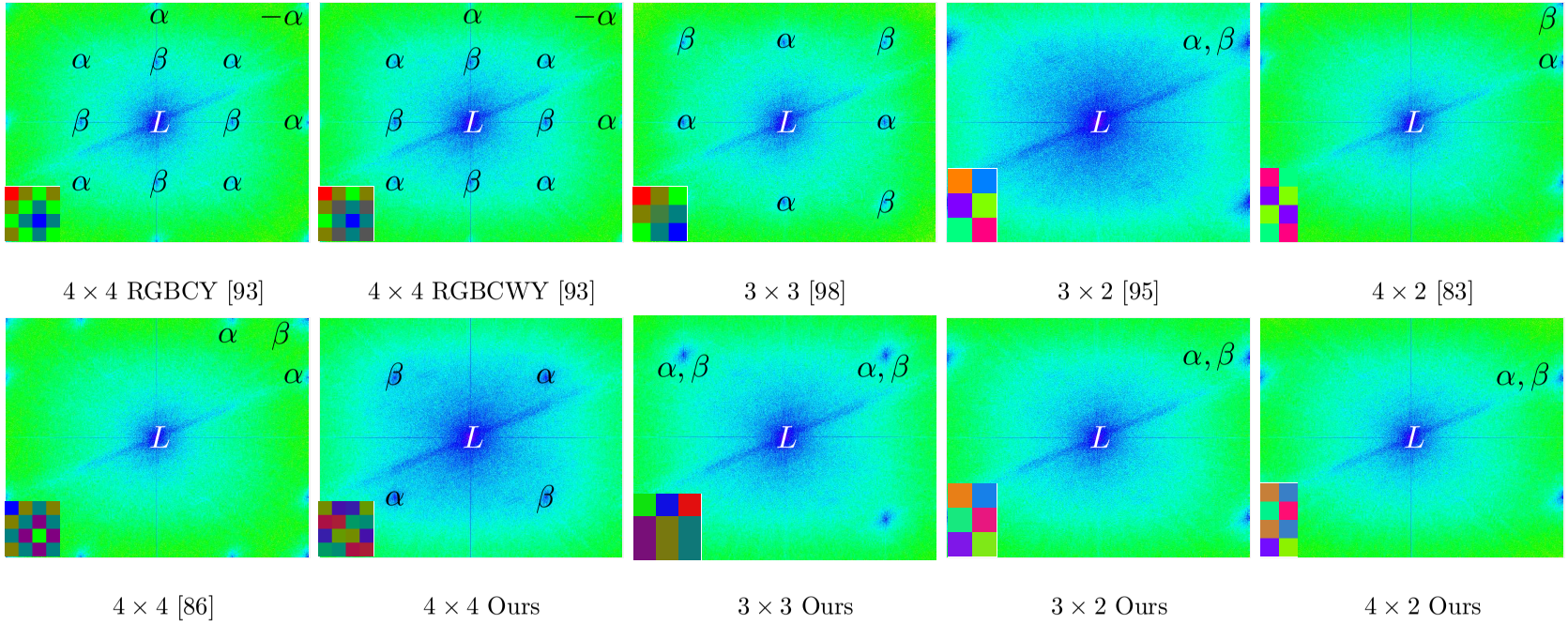
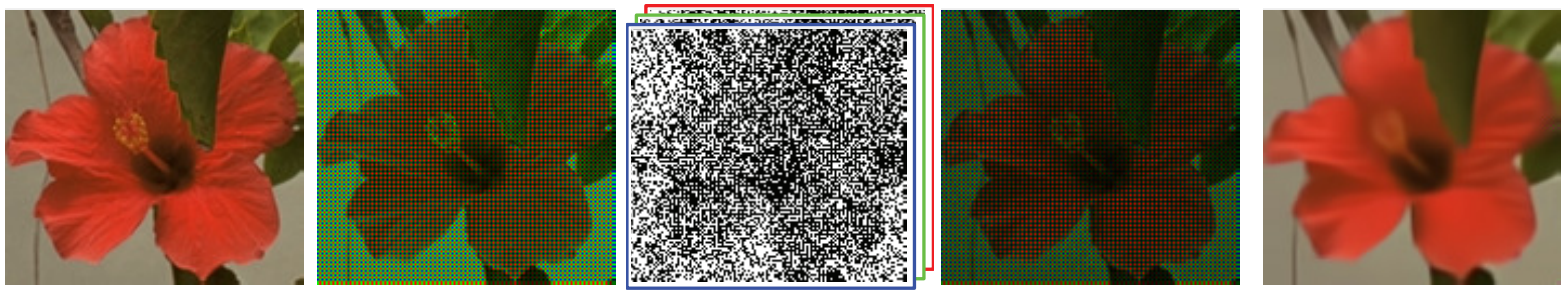


Fig. 6.9. Our proposed CFAs compared to other CFAs in literature. For every CFA, we show the spectrum of the "bike" Kodak image mosaicked by this CFA. We also show the organization of luminance (L) and chrominance channels (α and β) for CFAs that satisfy orthogonality constraint.

- 4×4: Among all 4×4 CFAs in Table B.1, [86] is the most robust CFA to aliasing, but the least robust to crosstalk; whereas RGBCWY [93] is the most robust to crosstalk, but it is not as robust to aliasing. Our CFA achieves the best of both worlds by having the same total-variation like RGBCWY, and good aliasing metric. Moreover, it has the lowest condition number.
- 3×3: Compared to [98], our CFA has less aliasing. The high aliasing metric of [98] is attributed to its design which overlooks frequency domain aliasing.
- 3×2: Compared to [95], our CFA is more robust to crosstalk and aliasing.
- 4×2: Compared to [83], our CFA has higher chrominance sensitivity and it is more robust to crosstalk.



(a) Ground Truth (b) Mosaicked Image (c) Binary Measurements (d) Summed Frame (e) Reconstruction, 31.64dB

Fig. 6.10. Simulation Setup: The ground truth image (a) is color filtered by a CFA to produce a mosaicked image (b). QIS generates $T = 1000$ binary frames (c) using the mosaicked image as light exposure. The T binary framed are summed to give an approximately clean image (d). Then, ADMM is applied to obtain the demosaicked image (e). Crosstalk is not added to this example, so there is no need for color correction.

6.7 Macbeth ColorChecker Reconstruction

In this experiment, we simulate the performance of different CFAs in reconstructing the Macbeth ColorChecker image. Pixel response is determined using the incident photon flux of D65 light and the spectral reflectance of Macbeth ColorChecker integrated over the visible light spectrum. QIS parameters and primary color filters are taken from [93]. For every CFA, we generate mosaicked images under two scenarios: 1) crosstalk kernels with leakage factors $(\alpha_r, \alpha_g, \alpha_b) = (0, 0, 0)$, i.e., no crosstalk, and 2) crosstalk kernels with leakage factors $(\alpha_r, \alpha_g, \alpha_b) = (0.45, 0.30, 0.20)$ as suggested by [93]. For fairness of comparison, we use Algorithm 3 for demosaicking all CFAs including RGBCY and RGBCWY CFAs proposed in [93]. Color correction with white balance is performed after color demosaicking for the crosstalk case for removing the crosstalk effect.

We use the following metrics [93] to evaluate the CFAs:

- Sensitivity metamerism index (SMI) which measures the drop in color reproduction accuracy due to crosstalk. It is obtained as a function of the CIEDE2000 color error metric which is obtained by calculating the mean square color difference in the CIELAB color space.

- Luminance SNR (YSNR) which measures the visual noise of luminance channel as defined in ISO 12232 [138].

Table B.1 shows these metrics for different CFAs with and without crosstalk. Our CFAs achieve higher color reproduction accuracy compared to others since they are optimized for crosstalk. This gain in color accuracy happens by trading the noise performance as observed by the drop of YPSNR metric.

6.8 Natural Color Image Reconstruction

In this experiment, we evaluate the performance of different CFAs for natural color image reconstruction. To this end, we use Kodak and McMaster color datasets

to generate 42 mosaicked images according to QIS model. QIS parameters are taken as $(q, \alpha, T) = (1, 2, 1000)$, and two scenarios are assumed: 1) No crosstalk, and 2) Moderate crosstalk with leakage factors $(\alpha_r, \alpha_g, \alpha_b) = (0.23, 0.15, 0.10)$. The low pass filter in Algorithm 3 is chosen as 21×21 Gaussian with standard deviation $\sigma = 21/3$ and multiplied by a hamming window to mitigate the windowing effect.

Table B.1 and shows the average color PSNR values on the 42 images. Our CFAs achieve better quality for the crosstalk case. Visually, our CFAs obtain color images with less aliasing artifacts and better details as shown in Figure B.3.

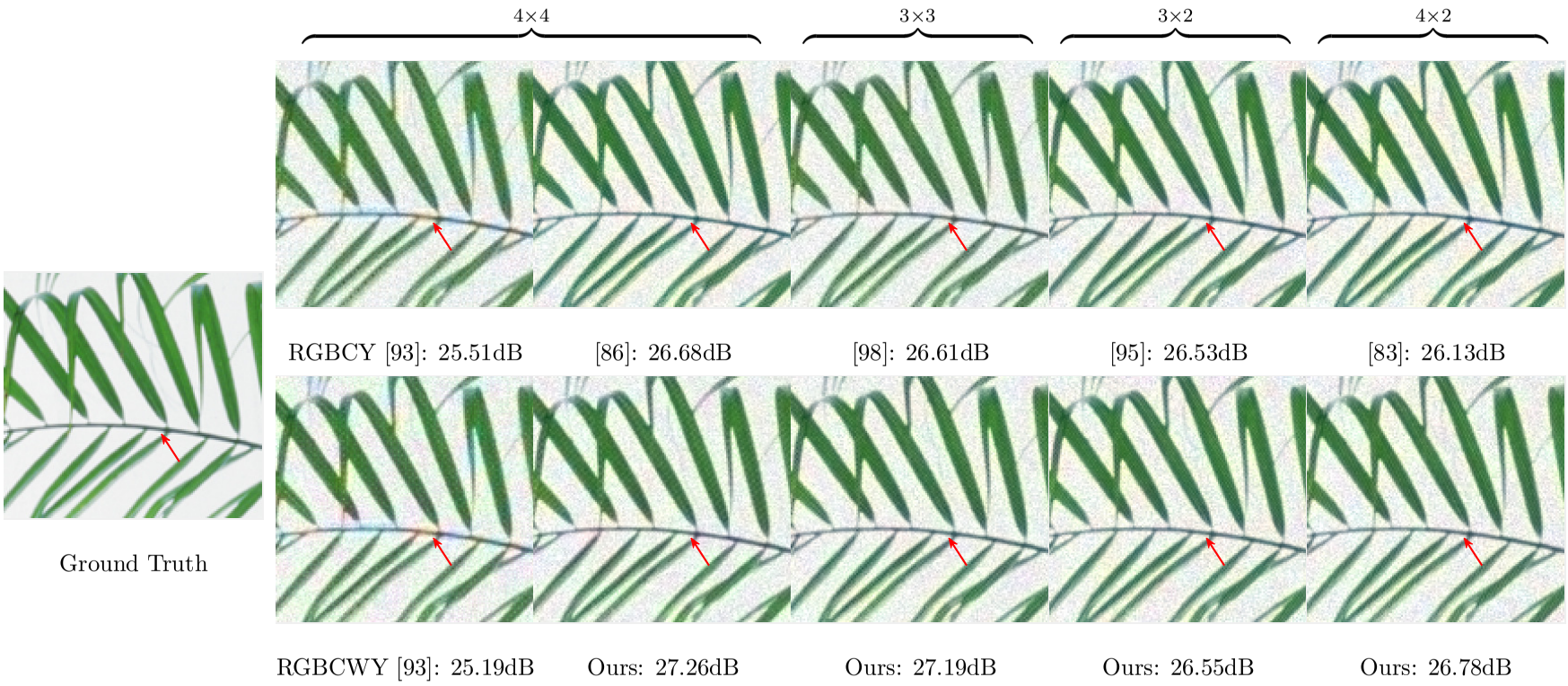


Fig. 6.11. Reconstructed color images from the QIS measurements. Each subfigure shows the result using a particular color filter array design.

7. CONCLUSION AND FUTURE DIRECTIONS

We studied three important problems related to QIS: 1) image reconstruction, 2) threshold design and 3) color filter array design. On the image reconstruction side, we proposed a non-iterative algorithms which can obtain a clean reconstruction with a significantly less computational complexity than existing work in literature. On the threshold design side, we proposed a practical threshold update scheme that can adapt the threshold to the incoming light both in space and time, i.e., it obtains a temporally-spatially-varying threshold. This scheme is based on a rigorous theoretical analysis for the reconstruction performance limits. As for the color filter array design, presented a general and flexible optimization framework to design color filter arrays for QIS. Our framework unifies the CMOS-based color filter array designs and extends to QIS. The color filter arrays designed by our framework are robust to crosstalk between the primary color channels, robust to aliasing between the luminance and chrominance channels, and are robust to noise. Experimental results show the effectiveness of our proposed methods compared to existing work in literature.

To achieve a practical and useful realization of QIS, several theoretical and technological issues require more exploration. The first challenge is to extend our QIS solutions to work with multi-bit QIS where the photon counts are quantized to a digital number with shallow bit-depth in the range $2-6$ bits. The second challenge is to obtain a fast QIS color image reconstruction scheme. Another important challenge is the handling of the binary data coming out from QIS. For a QIS with gigajots or more, read out at 1000 fps, the output data rate exceeds 1 Tb/s [28]. Efficient algorithms are crucial to handle this tremendous data rate efficiently. We discuss these issues in the next three sections in more details.

7.1 Extension to Multi-bit QIS

Multi-bit QIS uses an analog-to-digital (ADC) converter with shallow bit depth ($2-6$ bits) to quantize the photon counts. Compare to single-bit QIS, mutli-bit QIS allows capturing brighter scenes with the same integration time before saturating the pixel. However, this requires more complicated read-out circuit because the output data rate will be multiplied by the number of ADC bits.

On the signal processing side, QIS measurements still follow quantized Poisson process, but the quantization threshold in this case is $q = 2^n - 1$, where n is the number of ADC bits. Denote the QIS measurement of one jot by X, then we can write the distribution as follows:

$$X \sim \text{clip}\left(\text{Poisson}(\lambda), q\right), \tag{7.1}$$

where the probability mass function of X is defined as

$$p_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x < q \\ 1 - \Psi_q(\lambda), & x \geq q \end{cases} \tag{7.2}$$

For this random variable, we can compute its expectation as follows.

$$\mathbb{E}[X] = \sum_{x=0}^{q-1} x \frac{\lambda^x e^{-\lambda}}{x!} + q\left(1 - \Psi_q(\lambda)\right) \tag{7.3}$$

According to the value of n, we have two cases:

- Single-Bit QIS, $q = 1$: The expectation have a closed-form as we proved in Chapter 3:

  $$\mathbb{E}[X] = 1 - \Psi_q(\lambda) \tag{7.4}$$

  Thus, given a sample of T realizations of X: $\{x_1, \ldots, x_T\}$, the maximum likelihood estimate of the expectation is the sample mean since the distribution is Bernoulli, which is an exponential distribution [61].

  $$\frac{1}{T}\sum_{t=1}^{T} x_t \overset{\text{MLE}}{\approx} \mathbb{E}[X] = 1 - \Psi_q(\lambda) \tag{7.5}$$

Then, we can compute the latent light intensity λ by solving the equation $1 - \Psi_{q_{\max}}(\lambda) = \frac{1}{T}\sum_{t=1}^{T} x_t$ to get closed-form expression for λ

$$\lambda = \Psi_q^{-1}\left(1 - \frac{1}{T}\sum_{t=1}^{T} x_t\right) \tag{7.6}$$

In terms of image processing, this non-linear expression acts as a tone-mapping function that correct the contrast of the image obtained by averaging T QIS frames $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$.

- Multi-Bit QIS, $q \in \{3, 7, 15, 31, \ldots\}$: The expectation does not have a closed-form expression.

$$\mathbb{E}[X] = \sum_{x=0}^{q-1} x\frac{\lambda^x e^{-\lambda}}{x!} + q\left(1 - \Psi_q(\lambda)\right) \stackrel{\text{def}}{=} f(\lambda) \tag{7.7}$$

Given a sample of T realization of X, the sample mean is an approximation for the expectation; though, it is not the maximum likelihood estimate anymore because the distribution does not belong to the exponential family.

$$\frac{1}{T}\sum_{t=1}^{T} x_t \approx \mathbb{E}[X] = f(\lambda) \tag{7.8}$$

Then, we can get λ by applying the inverse function f^{-1} using a look-up table

$$\lambda = f^{-1}\left(\frac{1}{T}\sum_{i=1}^{T} x_i\right) \tag{7.9}$$

This inverse function acts as a tone mapping function. Without it (assuming $\mathbb{E}[X] = \lambda$), there is an error in the image contrast.

7.2 Fast Color QIS Image Reconstruction

One of the fastest methods for color image reconstruction is the demosaicking by frequency selection [73, 74, 78]. This method is fast because it requires simple multiplication operation for demodulation of chrominance channels and spatially-invariant filtering process for selecting specific channels at the baseband or passband.

In [95], the author tackles joint image denoising and demosaicking for images impaired with AWGN. He first reconstruct clean chrominance channels by demodulation to the baseband followed by low-pass filtering which is estimated in the least-square sense by minimizing the reconstruction error on a training dataset. After subtracting the re-modulated clean chrominance channels from the mosaicked image, the residual noise in the luminance channel is still AWGN. Hence, any off-the-shelf Gaussian denoiser can be used for luminance denoising.

For QIS, a similar approach to [95] can be used for fast reconstruction of color images. However, we should take care of the quantized Poisson noise model here, where there exist some open questions: Can we still find a better way to estimate or to learn the best low-pass filter for obtaining clean chrominance channels? After subtracting the chrominance channel from the mosaicked image, what is the noise model in the resulting luminance channel?

7.3 Handling the QIS Output Data

QIS outputs binary data at a huge data rate that can easily reach 1 Tb/sec. Transferring these binary measurements to an on-chip processor is not an easy task at all. The situation becomes worse if we do not quantize the photon counts because each photon count will need more than 1 bit to be represented, i.e., if each photon is represented by 4 bits, the data rate will be 1 Tb/sec. Nevertheless, this concern seems very legitimate because of this intuitive question: If QIS can count photons, why it throws away this invaluable information by an aggressive binary quantization? Photons are very valuable. However, QIS is forced to take this direction to decrease the output data rate.

A potential solution for decreasing the QIS data rate, in case of binary quantization, is to use *Source Coding*. Source coding is a well-established information theory problem which has been studied extensively for more than 60 years after the seminal work of Claude Shannon [139].

In the QIS case, we have cubicles of Bernoulli random variables with spatially variant Bernoulli parameter p, i.e., each cubicle of random variables has its own p which depends on the incoming light intensity on that cubicle. Specifically, p is related to the light exposure θ_m on the m-th jot according to (2.15) as follows: $p = 1-\Psi_q(\theta_m)$. In information theory literature, an important definition is the ϵ-typical set which is defined in the following proposition

Proposition 7.3.1 *For a sequence of n i.i.d. random variables $X_1, \ldots, X_n$ with probability mass function $p_X(x)$, the set of all sequences $(x_1, \ldots, x_n) \in \mathcal{R}_x^n$ such that*

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{7.10}$$

is called the ϵ-typical set $A_\epsilon^{(n)}$, where $H(X)$ is the entropy of the random variable X in bits.

A well-known result in information theory information theory is that the probability of the ϵ-typical set is close to one for sufficiently large n, i.e., $p(A_\epsilon^{(n)}) = 1 - \epsilon$. In other words, if we denote by $n = K^2T$ the number of jots in a QIS cubicle, as n increase, most probably we will observe the realizations $(x_1, \ldots, x_n)$ that belongs to $A_\epsilon^{(n)}$. This result is very useful because it means that we can only encode the sequences that belong to $A_\epsilon^{(n)}$, and ignore all other less probable sequences. This will decrease the number of bits required to represent the cubicle to be less than n (or equal to n in the worst case). The cardinality of $A_\epsilon^{(n)}$ is satisfies the following inequality:

$$|A_\epsilon^{(n)}| \leq 2^{n(H(x)+\epsilon)} \tag{7.11}$$

Hence, we need $n(H(x)+\epsilon)$ to represent the sequences in $A_\epsilon^{(n)}$. For a Bernoulli random variable with parameter p, the entropy $H(X)$ is calculated as

$$H(x) = -p\log_2(p) - (1-p)\log_2(1-p) \tag{7.12}$$

Figure 7.1 shows the variation of number of bits required to represent $A_\epsilon^{(n)}$. As p goes away from the point $p = 0.5$, we can achieve better compression. In QIS, p is

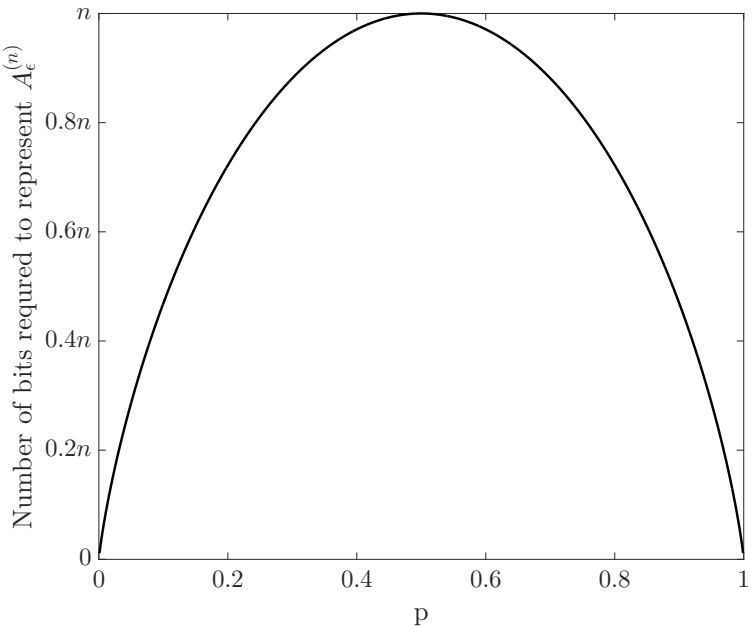


Fig. 7.1. Number of bits required to represent sequences $\{x_1, \ldots, x_n\}$ that belong to the ϵ-typical set, where ϵ is a sufficiently small positive number

obtained from the light intensity by this equation $p = 1 - \Psi_q(\theta)$. Therefore, if we have an initial estimate for the light intensity in each cubicle, we can efficiently compress the QIS data rate.

REFERENCES

REFERENCES

[1] J. Needham, *Science and Civilisation in China, vol.4, Physics and Physical Technology, Part 1, Physics*. New York, USA: Cambridge University Press, 1962.

[2] J. Campbell, *Film and Cinema Spectatorship: Melodrama and Mimesis, 1st ed*. Cambridge, UK: Polity Press, 2005.

[3] S. T. S. Al-Hassani, *1001 inventions: Muslim heritage in our world*, E. Woodcock and R. Saoud, Eds. Manchestar, UK: Foundation for Science, Technology and Civilisation, 2006.

[4] C. Sutton, "The impossibility of photography," *New Scientist*, pp. 40–43, December 1986, no. 1540 1541.

[5] R. Zakia and L. Stroebel, *The Focal Encyclopedia of Photography*, 3rd ed. Boston: Focal Press, 1993.

[6] F. Yang, "Bits from photons: Oversampled binary image acquisition," Ph.D. dissertation, École Polytechnique Fédérale De Lausanne, 2012.

[7] L. Day and I. McNeil, *Biographical Dictionary of the History of Technology*. Routledge, 1996.

[8] A. Einstein, "On a heuristic point of view concerning the production and transformation of light," *Annalen der Physik*, pp. 1–18, 1905.

[9] W. Boyle, "Nobel Lecture: CCD—An extension of man's view," *Review of Modern Physics*, vol. 82, pp. 2305–2306, August 2010. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.82.2305

[10] B. Hayes, "Computing science: Computational photography," *American Scientist*, vol. 96, no. 2, pp. 94–98, 2008.

[11] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH 2008 Classes*, ser. SIGGRAPH '08. New York, NY, USA: ACM, 2008, pp. 31:1–31:10. [Online]. Available: http://doi.acm.org/10.1145/1401132.1401174

[12] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *Proceedings of the 15th Pacific Conference on Computer Graphics and Applications (PG'07)*, Maui, HI, October 2007, pp. 382–390.

[13] K. Fife, A. El Gamal, and H. . P. Wong, "A multi-aperture image sensor with 0.7μmpixels in 0.11μm CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 12, pp. 2990–3005, Dec 2008.

[14] M. Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, August 2006.

[15] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, March 2008.

[16] E. R. Fossum, "What to do with sub-diffraction-limit (SDL) pixels?–A proposal for a gigapixel digital film sensor (DFS)," in *Proceedings of the 2005 IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors*, Karuizawa, Japan, September 2005, pp. 214–217.

[17] J. Ma, D. Hondongwa, and E. R. Fossum, "Jot devices and the quanta image sensor," in *Proceedings of the 2014 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, December 2014, pp. 10.1.1–10.1.4.

[18] J. Ma and E. R. Fossum, "A pump-gate jot device with high conversion gain for a quanta image sensor," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 2, pp. 73–77, March 2015.

[19] J. Ma, L. Anzagira, and E. R. Fossum, "A 1 μm-pitch quanta image sensor jot device with shared readout," *IEEE Journal of the Electron Devices Society*, vol. 4, no. 2, pp. 83–89, March 2016.

[20] F. Yang, Y. M. Lu, L. Sbaiz, and M. Vetterli, "An optimal algorithm for reconstructing images from binary measurements," in *Proceedings of the IS&T/SPIE Electronic Imaging Conference on Computational Imaging VIII*, vol. 7533, San Jose, CA, January 2010, pp. 75 330K–75 330K–12. [Online]. Available: http://dx.doi.org/10.1117/12.850887

[21] ——, "Bits from photons: Oversampled image acquisition using binary Poisson statistics," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1421–1436, April 2012.

[22] S. H. Chan and Y. M. Lu, "Efficient image reconstruction for gigapixel quantum image sensors," in *Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, December 2014, pp. 312–316.

[23] O. A. Elgendy and S. H. Chan, "Image reconstruction and threshold design for quanta image sensors," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP'16)*, Phoenix, AZ, September 2016, pp. 978–982.

[24] S. H. Chan, O. A. Elgendy, and X. Wang, "Images from bits: Non-iterative image reconstruction for quanta image sensors," *MDPI Sensors*, vol. 16, no. 11, November 2016, article number: 1961. [Online]. Available: http://www.mdpi.com/1424-8220/16/11/1961

[25] R. N. Clark, "Digital Camera Reviews and Sensor Performance Summary," http://www.clarkvision.com/articles/digital.sensor.performance.summary/, October 2016, accessed: 2019-04-15.

[26] E. R. Fossum, "Modeling the performance of single-bit and multi-bit quanta image sensors," *IEEE Journal of the Electron Devices Society*, vol. 1, no. 9, pp. 166–174, September 2013.

[27] N. Teranishi, "Required conditions for photon-counting image sensors," *IEEE Transactions on Electron Devices*, vol. 59, no. 8, pp. 2199–2205, August 2012.

[28] E. R. Fossum, J. Ma, and S. Masoodian, "Quanta image sensor: concepts and progress," in *Proceedings of the SPIE Commercial + Scientific Sensing and Imaging Conference on Advanced Photon Counting Techniques X*, vol. 9858, Baltimore, MD, May 2016, pp. 985 804–985 804–14. [Online]. Available: http://dx.doi.org/10.1117/12.2227179

[29] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, and E. Charbon, "Nonuniformity analysis of a 65k pixel CMOS SPAD imager," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 57–64, January 2016.

[30] N. A. W. Dutton, I. Gyongy, L. Parmesan, S. Gnecchi, N. Calder, B. R. Rae, S. Pellegrini, L. A. Grant, and R. K. Henderson, "A SPAD-based QVGA image sensor for single-photon counting and quanta imaging," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 189–196, January 2016.

[31] J. Hynecek, "Impactron-a new solid state image intensifier," *IEEE Transactions on Electron Devices*, vol. 48, no. 10, pp. 2238–2241, October 2001.

[32] M. S. Robbins and B. J. Hadwen, "The noise performance of electron multiplying charge-coupled devices," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1227–1232, May 2003.

[33] N. A. W. Dutton, I. Gyongy, L. Parmesan, and R. K. Henderson, "Single photon counting performance and noise analysis of CMOS SPAD-based image sensors," *MDPI Sensors*, vol. 16, no. 7, July 2016, article number: 1122. [Online]. Available: http://www.mdpi.com/1424-8220/16/7/1122

[34] B. F. Aull, D. R. Schuette, D. J. Young, D. M. Craig, B. J. Felton, and K. Warner, "A study of crosstalk in a 256×256 photon counting imager based on silicon Geiger-mode avalanche photodiodes," *IEEE Sensors Journal*, vol. 15, no. 4, pp. 2123–2132, April 2015.

[35] E. R. Fossum, "The Quanta Image Sensor (QIS): Concepts and Challenges," in *Proceedings of the OSA Technical Digest (CD), Optical Society of America*. Toronto, Canada: Optical Society of America, July 2011, paper JTuE1. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=ISA-2011-JTuE1

[36] J. Ma and E. R. Fossum, "Quanta image sensor jot with sub 0.3e- r.m.s. read noise and photon counting capability," *IEEE Electron Device Letters*, vol. 36, no. 9, pp. 926–928, September 2015.

[37] S. Masoodian, A. Rao, J. Ma, K. Odame, and E. R. Fossum, "A 2.5 pj/b binary image sensor as a pathfinder for quanta image sensors," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 100–105, January 2016.

[38] L. Sbaiz, F. Yang, E. Charbon, S. Susstrunk, and M. Vetterli, "The gigavision camera," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*, Taipei, Taiwan, April 2009, pp. 1093–1096.

[39] F. Yang, L. Sbaiz, E. Charbon, S. Süsstrunk, and M. Vetterli, "On pixel detection threshold in the gigavision camera," in *Proceedings of the IS&T/SPIE Electronic Imaging Conference on Digital Photography VI*, vol. 7537, San Jose, CA, January 2010, pp. 75 370G–75 370G–8. [Online]. Available: http://dx.doi.org/10.1117/12.840015

[40] N. A. W. Dutton, L. Parmesan, A. J. Holmes, L. A. Grant, and R. K. Henderson, "320 × 240 oversampled digital single photon counting image sensor," in *Proceedings of the 2014 Symposium on VLSI Circuits Digest of Technical Papers*, Honolulu, HI, June 2014, pp. 1–2.

[41] S. Burri, Y. Maruyama, X. Michalet, F. Regazzoni, C. Bruschini, and E. Charbon, "Architecture and applications of a high resolution gated SPAD image sensor," *Optics Express*, vol. 22, no. 14, pp. 17 573–17 589, July 2014. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-22-14-17573

[42] I. M. Antolovic, S. Burri, R. A. Hoebe, Y. Maruyama, C. Bruschini, and E. Charbon, "Photon-counting arrays for time-resolved imaging," *MDPI Sensors*, vol. 16, no. 7, June 2016, article number: 1005.

[43] T. Vogelsang and D. G. Stork, "High-dynamic-range binary pixel processing using non-destructive reads and variable oversampling and thresholds," in *Proceedings of the 2012 IEEE Sensors Conference*, Taipei, Taiwan, October 2012, pp. 1–4.

[44] T. Vogelsang, M. Guidash, and S. Xue, "Overcoming the full well capacity limit: high dynamic range imaging using multi-bit temporal oversampling and conditional reset," in *Proceedings of the 2013 International Image Sensor Workshop (IISW)*, Snowbird Resort, UT, June 2013.

[45] T. Vogelsang, D. G. Stork, and M. Guidash, "Hardware validated unified model of multibit temporally and spatially oversampled image sensors with conditional reset," *Journal of Electronic Imaging*, vol. 23, no. 1, p. 013021, February 2014. [Online]. Available: http://dx.doi.org/10.1117/1.JEI.23.1.013021

[46] "Andor ixon ultra 888 specifications," {http://www.andor.com/cameras/ixon-emccd-camera-series}, accessed: 2017-11-21.

[47] "Gigajot Technology LLC," http://www.gigajot.tech, accessed: 2019-04-09.

[48] S. Masoodian, J. M. D. Starkey, Y. Yamashita, and E. R. Fossum, "A 1mjot 1040fps 0.22e-rms stacked BSI quanta image sensor with cluster-parallel readout," in *Proceedings of the 2017 International Image Sensor Workshop (IISW)*, Hiroshima, Japan, May 2017, pp. 230–233.

[49] G. Grubbs, R. Michell, M. Samara, D. Hampton, and J.-M. Jahn, "A synthesis of star calibration techniques for ground-based narrowband electron-multiplying charge-coupled device imagers used in auroral photometry," *Journal*

of Geophysical Research: Space Physics, vol. 121, no. 6, pp. 5991–6002, 2016, 2015JA022186. [Online]. Available: http://dx.doi.org/10.1002/2015JA022186

[50] P. Seitz and A. J. Theuwissen, *Single-photon imaging.* Springer Science & Business Media, 2011, vol. 160.

[51] L. Liang, H. Shen, P. D. Camilli, and J. S. Duncan, "A novel multiple hypothesis based particle tracking method for clathrin mediated endocytosis analysis using fluorescence microscopy," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1844–1857, April 2014.

[52] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, "A fully digital 8 × 16 SiPM array for PET applications with per-pixel TDCs and real-time energy output," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 301–314, January 2014.

[53] S. P. Poland, N. Krstajić, J. Monypenny, S. Coelho, D. Tyndall, R. J. Walker, V. Devauges, J. Richardson, N. Dutton, P. Barber, D. D. Li, K. Suhling, T. Ng, R. K. Henderson, and S. M. Ameer-Beg, "A high speed multifocal multiphoton fluorescence lifetime imaging microscope for live-cell FRET imaging," *Biomedical Optics Express*, vol. 6, no. 2, pp. 277–296, February 2015. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-6-2-277

[54] I. Gyongy, T. A. Abbas, N. A. Dutton, and R. K. Henderson, "Object tracking and reconstruction with a quanta image sensor," in *Proceedings of the 2017 International Image Sensor Workshop (IISW)*, Hiroshima, Japan, May 2017, pp. 242–245, paper R22.

[55] L. J. Meng, "An intensified EMCCD camera for low energy Gamma ray imaging applications," *IEEE Transactions on Nuclear Science*, vol. 53, no. 4, pp. 2376–2384, August 2006.

[56] D. Shin, F. Xu, D. Venkatraman, R. Lussana, F. Villa, F. Zappa, V. K. Goyal, F. N. C. Wong, and J. H. Shapiro, "Photon-efficient imaging with a single-photon camera," *Nature Communications*, vol. 7, June 2016, article number: 12046. [Online]. Available: http://dx.doi.org/10.1038/ncomms12046

[57] E. Amri, Y. Felk, D. Stucki, J. Ma, and E. R. Fossum, "Quantum random number generation using a quanta image sensor," *MDPI Sensors*, vol. 16, no. 7, June 2016, article number: 1002.

[58] S. Masoodian, Y. Song, D. Hondongwa, J. Ma, K. Odame, and E. R. Fossum, "Early research progress on quanta image sensors," in *Proceedings of the 2013 International Image Sensor Workshop (IISW)*, Snowbird Resort, UT, June 2013.

[59] I. Gyongy, N. A. Dutton, L. Parmesan, A. Davies, R. Saleeb, R. Duncan, C. Rickman, P. Dalgarno, and R. K. Henderson, "Bit-plane processing techniques for low-light, high speed imaging with a spad-based qis," in *Proceedings of the 2015 International Image Sensor Workshop (IISW)*, Vaals, The Netherlands, June 2015, pp. 1–4.

[60] I. Gyongy, A. Davies, N. A. Dutton, R. Duncan, C. Rickman, R. K. Henderson, and P. Dalgarno, "Smart-aggregation imaging for single molecule localization with SPAD cameras," *Scientific Reports*, vol. 6, November 2016, article number: 37349.

[61] C. A. Bouman, "Model based image processing," 2013, [Online]. Available: https://engineering.purdue.edu/~bouman/publications/pdf/MBIP-book.pdf.

[62] C. Hu and Y. M. Lu, "Adaptive time-sequential binary sensing for high dynamic range imaging," in *Proceedings of the SPIE Defense, Security, and Sensing Conference on Advanced Photon Counting Techniques VI*, vol. 8375, Baltimore, MD, May 2012, pp. 83 750A–1. [Online]. Available: http://dx.doi.org/10.1117/12.919597

[63] F. Yang and M. Vetterli, "Oversampled noisy binary image sensor," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*, Vancouver, BC, Canada, May 2013, pp. 2060–2064.

[64] Y. M. Lu, "Adaptive sensing and inference for single-photon imaging," in *Proceedings of the 2013 47th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, March 2013, pp. 1–6.

[65] F. Yang, L. Sbaiz, E. Charbon, S. Süsstrunk, and M. Vetterli, "Image reconstruction in the gigavision camera," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, Kyoto, Japan, September 2009, pp. 2212–2219.

[66] T. Remez, O. Litany, and A. Bronstein, "A picture is worth a billion bits: Real-time image reconstruction from dense binary threshold pixels," in *Proceedings of the 2016 IEEE International Conference on Computational Photography (ICCP)*, Evanston, IL, May 2016, pp. 1–9.

[67] O. Litany, T. Remez, and A. Bronstein, "Image reconstruction from dense binary pixels," December 2015, [Online]. Available: http://arxiv.org/abs/1512.01774.

[68] E. R. Fossum, J. Ma, S. Masoodian, L. Anzagira, and R. Zizza, "The quanta image sensor: Every photon counts," *MDPI Sensors*, vol. 16, no. 8, August 2016, article number: 1260. [Online]. Available: http://www.mdpi.com/1424-8220/16/8/1260

[69] O. A. Elgendy and S. H. Chan, "Optimal threshold design for quanta image sensor," *IEEE Transactions on Computational Imaging*, vol. 4, no. 1, pp. 99–111, March 2018.

[70] A. Gnanasambandam, O. A. Elgendy, J. Ma, and S. H. Chan, "Megapixel photon-counting color imaging using quanta image sensor," March 2019, [Online]. Available: https://arxiv.org/abs/1903.09036.

[71] B. E. Bayer, "Color imaging array," USA Patent US3 971 065A, 1976.

[72] H. S. Malvar and R. Cutler, "High-quality linear interpolation for demosaicing of Bayer-patterned color images," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Montreal, Que., Canada, May 2004, pp. iii–485.

[73] D. Alleysson, S. Susstrunk, and J. Herault, "Linear demosaicing inspired by the human visual system," *IEEE Transactions on Image Processing*, vol. 14, no. 4, pp. 439–449, April 2005.

[74] E. Dubois, "Frequency-domain methods for demosaicking of Bayer-sampled color images," *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 847–850, Dec 2005.

[75] K. Hirakawa and T. W. Parks, "Adaptive homogeneity-directed demosaicing algorithm," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 360–369, March 2005.

[76] E. Dubois, "Filter Design for Adaptive Frequency-Domain Bayer Demosaicking," in *Proceedings of the 2006 International Conference on Image Processing (ICIP'06)*, Atlanta, GA, October 2006, pp. 2705–2708.

[77] B. Leung, G. Jeon, and E. Dubois, "Least-squares luma–chroma demultiplexing algorithm for Bayer demosaicking," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1885–1894, July 2011.

[78] G. Jeon and E. Dubois, "Demosaicking of noisy Bayer-sampled color images with least-squares luma-chroma demultiplexing and noise level estimation," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 146–156, Jan 2013.

[79] J. T. Korneliussen and K. Hirakawa, "Camera processing with chromatic aberration," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4539–4552, Oct 2014.

[80] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH Asia 2016*, vol. 35, no. 6, pp. 191:1–191:12, Nov. 2016. [Online]. Available: http://doi.acm.org/10.1145/2980179.2982399

[81] H. Tan, X. Zeng, S. Lai, Y. Liu, and M. Zhang, "Joint demosaicing and denoising of noisy Bayer images with ADMM," in *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP'17)*, Beijing, China, September 2017, pp. 2951–2955.

[82] R. Lukac and K. N. Plataniotis, "Color filter arrays: design and performance analysis," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 4, pp. 1260–1267, November 2005.

[83] K. Hirakawa and P. J. Wolfe, "Spatio-spectral color filter array design for optimal image recovery," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1876–1890, Oct 2008.

[84] Y. M. Lu and M. Vetterli, "Optimal color filter array design: quantitative conditions and an efficient search procedure," in *Proceedings of the IS&T/SPIE Electronic Imaging Conference on Digital Photography V*, vol. 7250, San Jose, CA, January 2009, pp. 7250 – 7250 – 8. [Online]. Available: https://doi.org/10.1117/12.807598

[85] L. Condat, "Color filter array design using random patterns with blue noise chromatic spectra," *Image and Vision Computing*, vol. 28, no. 8, pp. 1196 – 1202, August 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885609002741

[86] P. Hao, Y. Li, Z. Lin, and E. Dubois, "A geometric method for optimal design of color filter arrays," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 709–722, March 2011.

[87] J. Wang, C. Zhang, and P. Hao, "New color filter arrays of high light sensitivity and high demosaicking performance," in *Proceedings of the 2011 IEEE International Conference on Image Processing (ICIP'11)*, Brussels, Belgium, September 2011, pp. 3153–3156.

[88] A. Chakrabarti, W. T. Freeman, and T. Zickler, "Rethinking color cameras," in *Proceedings of the 2014 IEEE International Conference on Computational Photography (ICCP)*, Santa Clara, CA, May 2014, pp. 1–8.

[89] P. Amba, J. Dias, and D. Alleysson, "Random color filter arrays are better than regular ones," *Color and Imaging Conference*, vol. 2016, no. 1, pp. 294–299, 2016. [Online]. Available: https://www.ingentaconnect.com/content/ist/cic/2016/00002016/00000001/art00052

[90] C. Bai, J. Li, Z. Lin, and J. Yu, "Automatic design of color filter arrays in the frequency domain," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1793–1807, April 2016.

[91] J. Li, C. Bai, Z. Lin, and J. Yu, "Automatic design of high-sensitivity color filter arrays with panchromatic pixels," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 870–883, Feb 2017.

[92] J. Li, C. Bai, Z. Lin, and J. Yu, "Optimized color filter arrays for sparse representation-based demosaicking," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2381–2393, May 2017.

[93] L. Anzagira and E. R. Fossum, "Color filter array patterns for small-pixel image sensors with substantial cross talk," *Journal of the Optical Society of America A*, vol. 32, no. 1, pp. 28–34, January 2015. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-32-1-28

[94] C. Chao, H.-Y. Tu, K.-Y. Chou, P.-S. Chou, F.-L. Hsueh, V. Wei, R.-J. Lin, and B.-C. Hseih, "Crosstalk metrics and the characterization of 1.1 μm-pixel cis," in *Proceedings of International Image Sensor Workshop (IISW)*, Hokkaido, Japan, June 2011, p. R7.

[95] L. Condat, "A new color filter array with optimal properties for noiseless and noisy color image acquisition," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2200–2210, Aug 2011.

[96] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3081–3089. [Online]. Available: http://papers.nips.cc/paper/6251-learning-sensor-multiplexing-design-through-back-propagation.pdf

[97] B. Henz, E. S. L. Gastal, and M. M. Oliveira, "Deep joint design of color filter arrays and demosaicing," *Computer Graphics Forum*, vol. 37, no. 2, pp. 389–399, May 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13370

[98] H. Biay-Cheng, H. Siddiqui, J. Luo, G. Todor, and A. Kalin, "New color filter patterns and demosaic for sub-micron pixel arrays," 2015.

[99] H. Siddiqui, K. Atanassov, and S. Goma, "Hardware-friendly universal demosaick using non-iterative map reconstruction," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP'16)*, Phoenix, AZ, Sept 2016, pp. 1794–1798.

[100] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Courier Corporation, 1964, no. 55.

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

[102] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, December 2006.

[103] L. Zhang and W. Zuo, "Image restoration: From sparse and low-rank priors to deep priors [lecture notes]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 172–179, September 2017.

[104] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-Play priors for model based reconstruction," in *Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP'13)*, Austin, TX, December 2013, pp. 945–948.

[105] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, August 2007.

[106] S. H. Chan, T. Zickler, and Y. M. Lu, "Monte Carlo non-local means: Random sampling for large-scale image filtering," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3711–3725, August 2014.

[107] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, July 2017.

[108] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, September 2017.

[109] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajkak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, J. Kautz, and K. Pulli, "FlexISP: A flexible camera image processing framework," *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH Asia 2014*, vol. 33, no. 6, pp. 231:1–231:13, November 2014. [Online]. Available: http://doi.acm.org/10.1145/2661229.2661260

[110] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 84–98, March 2017.

[111] M. Makitalo and A. Foi, "Optimal inversion of the generalized anscombe transformation for Poisson - Gaussian noise," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 91–103, January 2013.

[112] L. Azzari and A. Foi, "Variance stabilization for noisy+estimate combination in iterative Poisson denoising," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1086–1090, August 2016.

[113] A. Foi, "Clipped noisy images: Heteroskedastic modeling and practical denoising," *Signal Processing*, vol. 89, no. 12, pp. 2609–2629, December 2009.

[114] J. Salmon, Z. Harmany, C. Deledalle, and R. Willet, "Poisson noise reduction with non-local PCA," *Journal of Mathematical Imaging and Vision*, vol. 48, no. 2, pp. 279–294, February 2014.

[115] Z. T. Harmany, R. F. Marcia, and R. M. Willet, "This is SPIRAL-TAP: sparse Poisson intensity reconstruction algorithms: Theory and practice," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1084–1096, September 2011.

[116] A. Rond, R. Giryes, and M. Elad, "Poisson inverse problems by the Plug-and-Play scheme," *Journal of Visual Communication and Image Representation*, vol. 41, no. Supplement C, pp. 96–108, September 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320316301985

[117] A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*. Pearson Prentice Hall, 2008.

[118] L. Wasserman, *All of nonparametric statistics*. New York, USA: Springer-Verlag, 2006.

[119] F. J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data," *Biometrika*, vol. 35, no. 3-4, pp. 246–254, 1948.

[120] L. Brown, T. Cai, and A. DasGupta, "On selecting a transformation : with applications," [Online]. Available: http://www.stat.purdue.edu/~dasgupta/vst.pdf.

[121] J. Ma, D. Starkey, A. Rao, K. Odame, and E. R. Fossum, "Characterization of quanta image sensor pump-gate jots with deep sub-electron read noise," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 6, pp. 472–480, November 2015.

[122] E. R. Fossum, "Multi-bit quanta image sensors," in *Proceedings of the 2015 International Image Sensor Workshop (IISW)*, Vaals, The Netherlands, June 2015, pp. 292–295.

[123] I. Sprow, D. Kuepper, Z. Barańczuk, and P. Zolliker, "Image quality assessment using a high dynamic range display," in *Proceedings of the 12th Congress of the International Colour Association*, Newcastle Gateshead, UK, July 2013, p. 307–310.

[124] S. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH Asia 2016*, vol. 35, no. 6, pp. 192:1–192:12, November 2016. [Online]. Available: http://www.hdrplusdata.org/hdrplus.pdf

[125] K. I. Schultz, M. W. Kelly, J. J. Baker, M. H. Blackwell, M. G. Brown, C. B. Colonero, C. L. David, B. M. Tyrrell, and J. R. Wey, "Digital-pixel focal plane array technology," *Lincoln Laboratory Journal*, vol. 20, no. 2, pp. 36–51, 2014.

[126] Y. Hel-Or, "The canonical correlations of color images and their use for demosaicing," *HP Laboratories Israel, Tech. Rep. HPL-2003-164R1*, 2004.

[127] K. Hirakawa, "Cross-talk explained," in *Proceedings of the 2008 15th IEEE International Conference on Image Processing (ICIP'08)*, San Diego, CA, October 2008, pp. 677–680.

[128] Z. Opial, "Weak convergence of the sequence of successive approximations for nonexpansive mappings," *Bulletin of the American Mathematical Society*, vol. 73, no. 4, pp. 591–597, 1967.

[129] M. Stein, "Large sample properties of simulations using latin hypercube sampling," *Technometrics*, vol. 29, no. 2, pp. 143–151, 1987. [Online]. Available: https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1987.10488205

[130] T. Yap-Peng and A. Tinku, "Method for color correction with noise consideration," in *Proceedings of the SPIE Electronic Imaging Conference on Color Imaging: Device-Independent Color, Color Hardcopy, and Graphic Arts V*, vol. 3963, San Jose, CA, December 1999, pp. 3963 – 3963 – 9. [Online]. Available: https://doi.org/10.1117/12.373413

[131] "HDR-eye dataset," http://mmspg.epfl.ch/hdr-eye, accessed: 2019-04-09.

[132] H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Visual attention in LDR and HDR images," in *Proceedings of the 9th International Workshop on Video Process. and Quality Metrics for Consumer Electronics (VPQM)*, Chandler, AZ, February 2015.

[133] K. Zhang, W. Zuo, G. Wangmeng, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proceedings of the IEEE 2017 Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017, pp. 3929–3938.

[134] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, Hilton Head Island, SC, June 2000, pp. 472–479 vol.1.

[135] C. Aguerrebere, A. Almansa, J. Delon, Y. Gousseau, and P. Muse, "A Bayesian hyperprior approach for joint image denoising and interpolation, with an application to HDR imaging," *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 633–646, December 2017.

[136] "Kodak color dataset," http://r0k.us/graphics/kodak/, accessed: 2019-04-12.

[137] Z. Lei, W. Xiaolin, B. Antoni, and L. Xin, "Color demosaicking by local directional interpolation and nonlocal adaptive thresholding," *J. Electron. Imaging*, vol. 20, no. 2, pp. 1 – 17 – 17, April 2011. [Online]. Available: https://doi.org/10.1117/1.3600632

[138] ISO, "Photography–digital still cameras–determination of exposure index, ISO speed ratings, standard output sensitivity, and recommended exposure index," ISO 12232:2006 (International Organization for Standardization), Geneva, Switzerland, 2006.

[139] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

APPENDICES

A. SUPPLEMENTARY MATERIAL FOR CHAPTER 4

This supplementary appendix provides the following additional information for Chapter 4

- Derivation of $\mathrm{SNR}_q(c)$ from exposure-referred SNR,
- Properties of the incomplete Gamma function,
- Comparison with the threshold design scheme by Yang [6],
- Phase transition under different configurations,
- Influence of Non-Boxcar Kernel $\mathbf{G}$, and
- Additional results for HDR image reconstruction.

A.1 Derivation of $\mathrm{SNR}_q(c)$ from exposure-referred SNR

In the literature of QIS devices, one metric to quantify the image quality is the *exposure*-referred signal-to-noise [26]. In image processing, however, exposure-referred SNR is not commonly used. The goal of this section is to show that the SNR we showed in the main article is equivalent to the exposure-referred SNR.

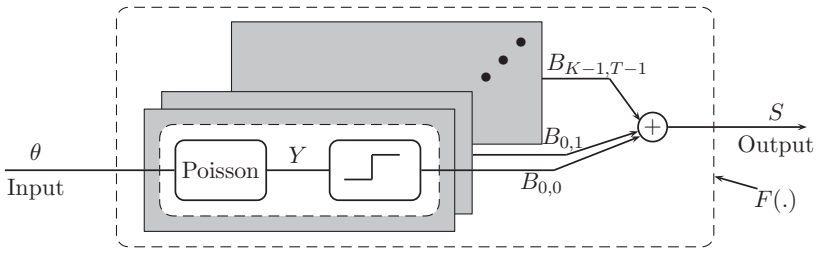


Fig. A.1. Block diagram illustrating a QIS with input-output relation output $= F(\text{input})$

To understand the exposure-referred SNR, we have to first understand two common ways of defining a signal to noise ratio. Consider the truncated Poisson part of the QIS model shown in Figure A.1. The input to this model is the over-sampled measurement θ. The truncated Poisson process can be considered as a black box function F which takes an input θ and generates an output S, defined as

$$S = \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} B_{k,t}, \tag{A.1}$$

where $\mathcal{B}_n = \{B_{k,t} \mid k = 0, 1, \ldots, K-1,\ t = 0, 1, \ldots, T-1\}$ is the spatial-temporal block containing all binary bits corresponding to θ. As shown in the main article, the mean and variance of S are

$$\mathbb{E}[S] = KT(1 - \Psi_q(\theta)), \quad \mathrm{Var}[S] = KT\Psi_q(\theta)(1 - \Psi_q(\theta)), \tag{A.2}$$

respectively.

The first notion of signal-to-noise, which is the one used in CCD and CMOS, is called the output-referred SNR. $\mathrm{SNR}_{\mathrm{OR}}$ is defined as the ratio between the output signal and the photon shot noise. Referring to Figure A.1, this is

$$\mathrm{SNR}_{\mathrm{OR}} = \frac{\text{output signal}}{\text{noise}} = \frac{\mathbb{E}[S]}{\sqrt{\mathrm{Var}[S]}} = \sqrt{KT\frac{1 - \Psi_q(\theta)}{\Psi_q(\theta)}}. \tag{A.3}$$

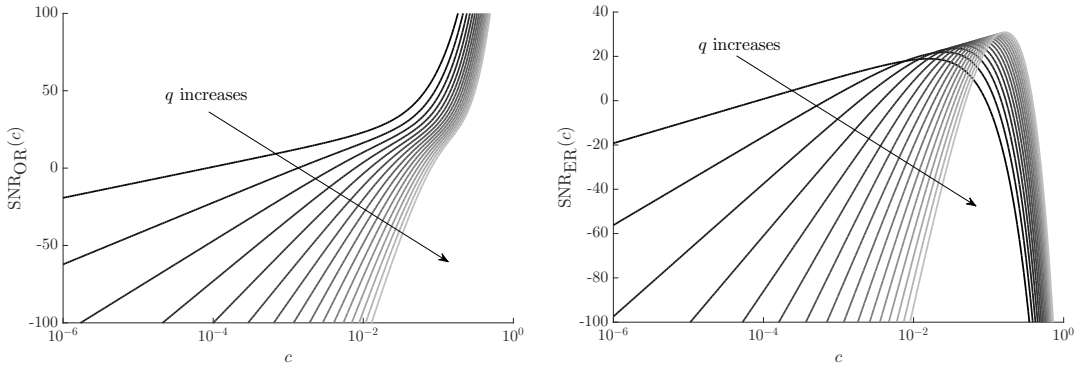
However, $\mathrm{SNR}_{\mathrm{OR}}$ fails to work for QIS because the shot noise is arbitrarily small if all bits are 1 or 0. In [26], Fossum called it squeezing of the noise. If we plot $\mathrm{SNR}_{\mathrm{OR}}$ as a function of θ, then we observe that $\mathrm{SNR}_{\mathrm{OR}}$ approaches to infinity as θ grows.

The second notion of signal-to-noise, which is a modification of $\mathrm{SNR}_{\mathrm{OR}}$, is the exposure-referred SNR. $\mathrm{SNR}_{\mathrm{ER}}$ is the ratio between the exposure signal θ and the exposure-referred noise. This noise is defined as [26]:

$$\text{Exposure-referred noise} = \frac{d\theta}{d\mathbb{E}[S]}\sqrt{\mathrm{Var}[S]}$$

The factor $\frac{d\theta}{d\mathbb{E}[S]}$ can be considered as the "inverse" transfer function from the output to the input. $\frac{d\theta}{d\mathbb{E}[S]}$ can be determined by taking derivative of the expectation in (A.2) with respect to $\mathbb{E}[S]$

$$\frac{d\mathbb{E}[S]}{d\mathbb{E}[S]} = \frac{dKT\left(1 - \Psi_q(\theta)\right)}{d\mathbb{E}[S]}$$



(a) $\mathrm{SNR}_{\mathrm{OR}}$ (b) $\mathrm{SNR}_{\mathrm{ER}}$

Fig. A.2. Comparison of the SNRs for $q \in \{1, \ldots, 16\}$. In this experiment, we fix $\alpha = 400$, $K = 4$, and $T = 30$.

Using chain rule, we observe that

$$1 = -KT \frac{d}{d\theta} \Psi_q(\theta) \frac{d\theta}{d\mathbb{E}[S]}$$

Since $\frac{d}{d\theta}\Psi_q(\theta) = \frac{-e^{-\theta}\theta^{q-1}}{\Gamma(q)}$, it holds that

$$1 = -KT \left(\frac{-e^{-\theta}\theta^{q-1}}{\Gamma(q)} \right) \frac{d\theta}{d\mathbb{E}[S]}$$

Hence,

$$\frac{d\theta}{d\mathbb{E}[S]} = \frac{\Gamma(q)}{KTe^{-\theta}\theta^{q-1}}$$

The exposure-referred SNR is defined as

$$\begin{aligned} \mathrm{SNR}_{\mathrm{ER}} &= \frac{\text{exposure signal}}{\text{exposure-referred noise}} \\ &= \frac{\theta}{\sqrt{\mathrm{Var}[S]}\frac{d\theta}{d\mathbb{E}[S]}} \\ &= \frac{e^{-\theta}\theta^{q}}{\Gamma(q)} \sqrt{\frac{KT}{\Psi_q(\theta)\big(1 - \Psi_q(\theta)\big)}}. \end{aligned}$$

Taking logarithm shows that $\mathrm{SNR}_{\mathrm{ER}}$ is identical to the SNR derived from the Fisher Information shown in the main article.

A.2 Properties of the incomplete Gamma function

In the main article, we used the incomplete Gamma function for QIS analysis. In this section, we provide more details about the properties of the incomplete Gamma function.

First, we recall that the normalized upper incomplete Gamma function $\Psi_q : \mathbb{R}^+ \to [0, 1]$ is defined as

$$\Psi_q(\theta) \stackrel{\text{def}}{=} \frac{1}{\Gamma(q)} \int_\theta^\infty t^{q-1} e^{-t} dt, \quad \text{for } \theta > 0,\ q \in \mathbb{N}. \tag{A.4}$$

where $\Gamma(q) = (q-1)!$ is the standard Gamma function.

In this equation, we note that $\Psi_q(\theta)$ depends on two variables: q and θ.

- As a function of θ. As we showed in the main article, $\Psi_q(\theta)$ is a monotonically decreasing function of θ because the derivative is negative:

$$\frac{d}{d\theta}\Psi_q(\theta) = \frac{-\theta^{q-1}e^{-\theta}}{\Gamma(q)} < 0.$$

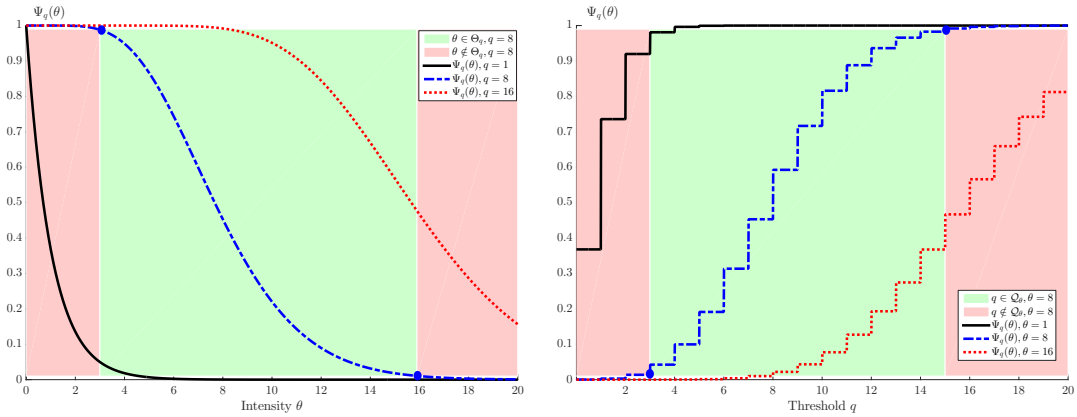
  However, $\Psi_q(\theta)$ is very close to 1 when θ is small, and is very close to 0 when θ is large. Therefore, there exists a range of θ in which $\Psi_q(\theta)$ can attain a reasonably good inverse. We define this set as the θ-admissible set

$$\Theta_q \stackrel{\text{def}}{=} \{\theta \mid \varepsilon \leq \Psi_q(\theta) \leq 1 - \varepsilon\}, \tag{A.5}$$

  for any fixed q and a tolerance level ε. An illustration of Θ_q is shown in Figure A.3.

- As a function of q. The incomplete Gamma function $\Psi_q(\theta)$ can also be considered as a function of q. In this case, $\Psi_q(\theta)$ is only defined for integer values of q. We illustrate the behavior of $\Psi_q(\theta)$ as a function of q in Figure A.3. The set of q in which $\Psi_q(\theta)$ is sufficiently away from 0 and 1 is defined as the q-admissible set.

$$\mathcal{Q}_\theta \stackrel{\text{def}}{=} \{q \mid \varepsilon \leq \Psi_q(\theta) \leq 1 - \varepsilon\}. \tag{A.6}$$



(a) $\Psi_q(\theta)$ vs. θ (b) $\Psi_q(\theta)$ vs. q

Fig. A.3. $\Psi_q(\theta)$ as a function of θ and q. In defining, $\mathcal{Q}_\theta$ and Θ_q, we set $\epsilon = 0.01$.

A.3 Comparison with the threshold design scheme by Yang [6]

In this section, we compare our threshold scheme with the one in [6].

First, we recall that the optimality of our method is based on a lower-bound $L_q(c)$ for the per-pixel SNR:

$$q^*(c) = \underset{q\in\mathbb{N}}{\text{argmax}}\ \text{SNR}_q(c) \approx \underset{q\in\mathbb{N}}{\text{argmax}}\ L_q(c) = \left\lfloor \frac{\alpha c}{K} \right\rfloor \tag{A.7}$$

Therefore, the optimal threshold is a function of c, which changes in space and in time.

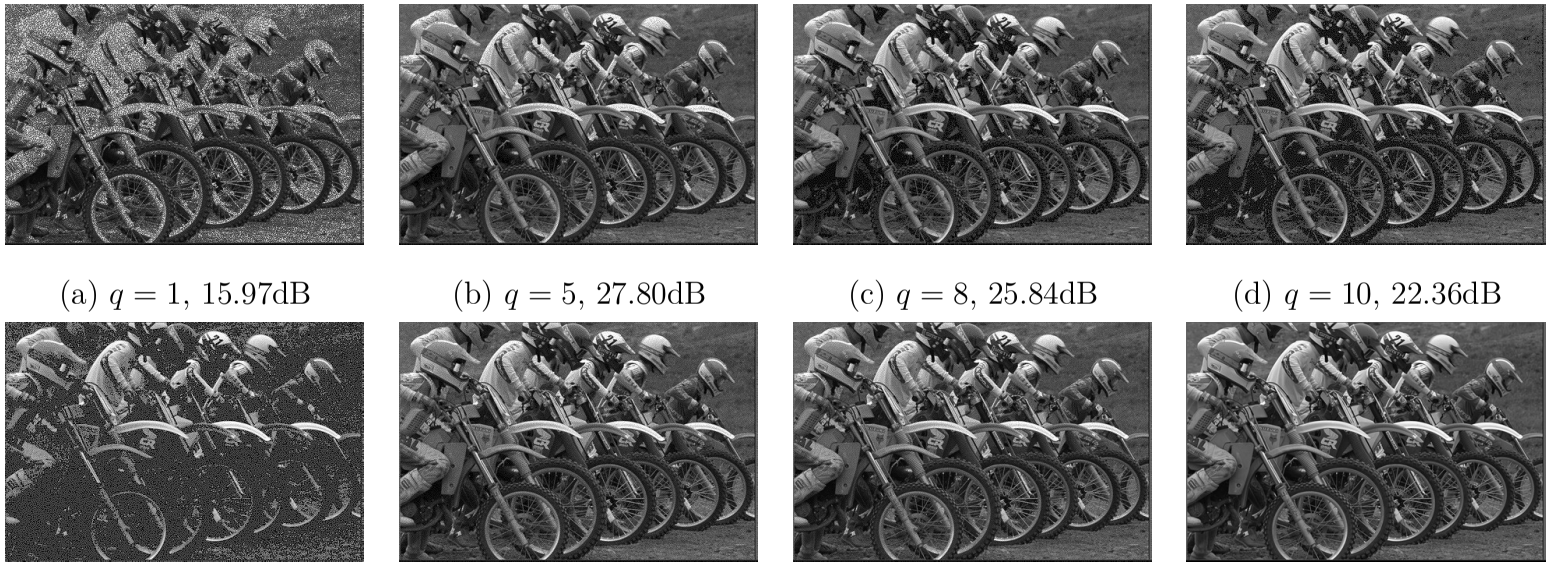
In contrast, [6] uses a checkerboard pattern by alternating two thresholds (q_1^*, q_2^*). These two thresholds are obtained by maximizing the Cramér-Rao lower bound (CRLB) over a range of light intensity values $[c_{\min}, c_{\max}]$:

$$(q_1^*, q_2^*) = \underset{1\le q_1, q_2 \le q_{\max}}{\text{argmin}} \int_{c_{\min}}^{c_{\max}} CRLB(q_1, q_2, c)\, dc. \tag{A.8}$$

As a result, the threshold is optimal in the *average sense*. To compare the two approaches, we followed the same steps in [6] to obtain $CRLB(q_1, q_2, c)$ for a checkerboard pattern in terms of $\Psi_q(c)$ as follows

$$CRLB(q_1, q_2, c) = \sum_{i=1}^{2} \frac{\alpha^2}{2K} \frac{e^{-2\theta}\theta^{(2(q_i-1))}}{\Gamma(q_i)^2 \Psi_{q_i}(\theta)\left[1 - \Psi_{q_i}(\theta)\right]} \tag{A.9}$$

where $\theta = \alpha c/K$. Using the parameters $\alpha = K(q_{\max -1})$, $q_{\max} = 16$, $K = 4$, and using trapezoidal technique for numerical integration over c, we obtain that $q_1^* = 4$ and $q_2^* = 12$. Figure A.4 shows the reconstructed images using uniform threshold maps with thresholds $q \in \{1, 5, 8, 10, 15\}$, the checkerboard threshold map in [6] with $q_1^* = 4$ and $q_2^* = 12$, and the oracle threshold map obtained by (A.7). In this experiment, our proposed method achieves 28.15 dB, which is 0.83 dB higher than the checkerboard pattern.



(a) $q = 1$, 15.97dB (b) $q = 5$, 27.80dB (c) $q = 8$, 25.84dB (d) $q = 10$, 22.36dB

(e) $q = 15$, 15.35dB (f) [6], 27.32dB (g) Proposed, 28.15dB (h) Ground Truth

Fig. A.4. Spatial oversampling $K = 4$. Temporal oversampling $T = 20$. Quadratic B-spline kernel is used in synthesis and reconstruction models. Gradient descent is used to obtain the ML estimate. For bisection threshold map, 8 frames are used for adapting the map, and 12 frames are used for reconstruction. For all other maps, the whole 20 frames are used for reconstruction.

A.4 Phase transition under different configurations

In the main article, we showed the phase transition behavior of the ML estimate using $K = 4$, $T = 50$, and $\delta = 2 \times 10^{-4}$. In this section, we study the effect of changing K, T, and δ on the phase transition region width.

As a function of T. Figure A.5-Figure A.6 illustrate the phase transition behavior when $T = 10, 25, 50$, and 100. As T increases, the width of the green region increases. However, if we fix the range of the bit density $1 - \mathbb{E}[\gamma_q(c)]$, we observe that the SNR does not vary significantly even as T changes.

As a function of K. The spatial oversampling K affects both the threshold $q^*(c) = \lfloor \alpha c/K \rfloor + 1$ and the phase transition width. Figure A.7(a) illustrates the behavior of the threshold q^* as a function of K. As K increases, q^* decreases. However, the optimal q^* still stays within the set $\mathcal{Q}_\theta$.

As a function of δ. The constant δ is used to define the set $\mathcal{Q}_\theta$:

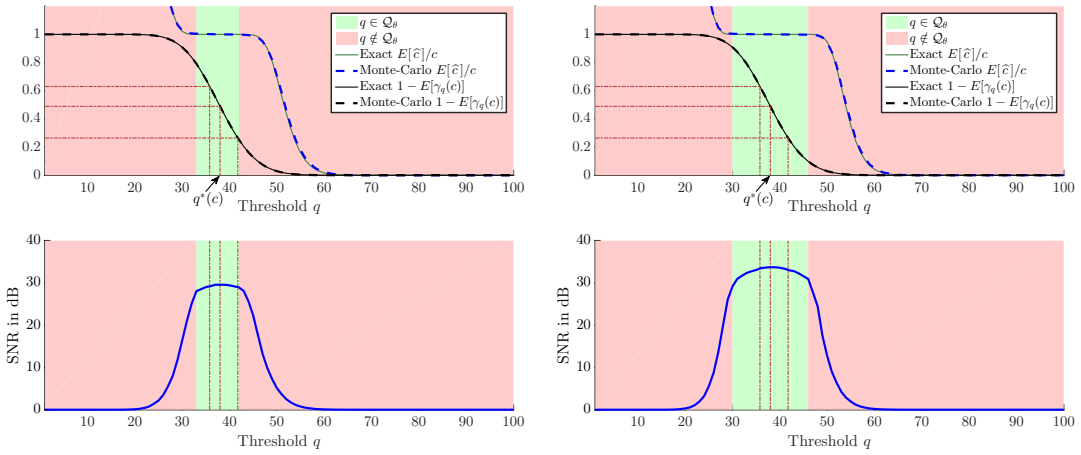
$$\mathcal{Q}_\theta \stackrel{\text{def}}{=} \left\{ q \;\middle|\; 1 - \left(\frac{\delta}{2}\right)^{\frac{1}{KT}} \leq \Psi_q(\theta) \leq \left(\frac{\delta}{2}\right)^{\frac{1}{KT}} \right\}. \tag{A.10}$$

The constant δ is the tolerance level. When δ increases, the size of the set $\mathcal{Q}_\theta$ should also increase. This result is shown in Figure A.7(b).

Using the closed form expression of the average bit density $1 - \Psi_q(\theta)$, we can calculate the average bit density at the optimal threshold $q^* = \lfloor \theta \rfloor + 1$, which is shown in Figure A.8. We notice that as long as $\theta \geq 1$, the average bit density is between 0.264 and 0.630. Within this range, we observe from Figure A.5-Figure A.6 that the SNR does not vary significantly if the estimated threshold is deviated from the optimal threshold. This observation relaxes the requirement of the bisection method from obtaining the exact optimal threshold to obtaining a threshold that make the bit density equal to 0.5. Since $0.5 \in [0.264, 0.630]$, we guarantee to achieve an SNR which is sufficiently close to the optimal SNR.

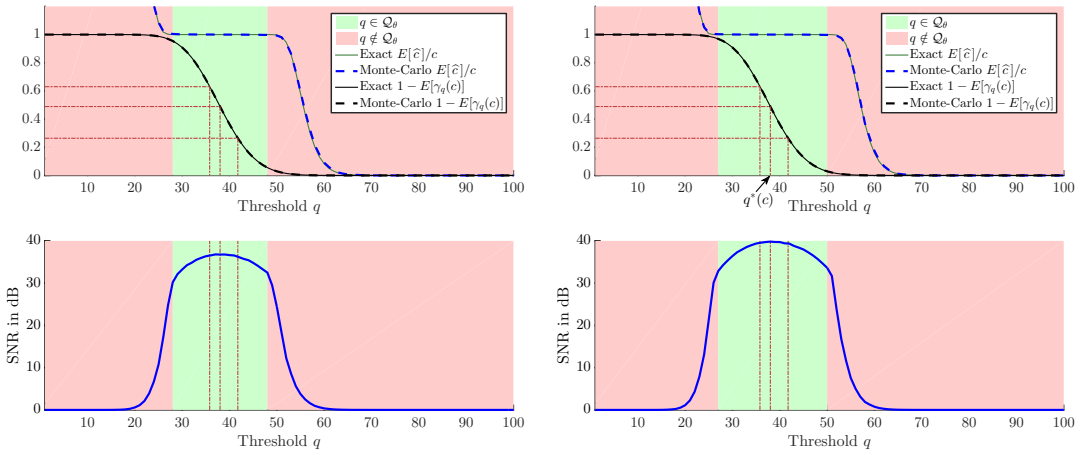
Controlling $\theta \geq 1$ can be achieved by tuning the constant α. Tuning α can be hardware-implemented by increasing the exposure period. Intuitively what $\theta \geq 1$ requires is that the average number of impinging photons per jot must be at least

one. If θ is less than one, then most bits will become zeros. Increasing exposure period (i.e., increasing α) will ensure sufficient number of photons.



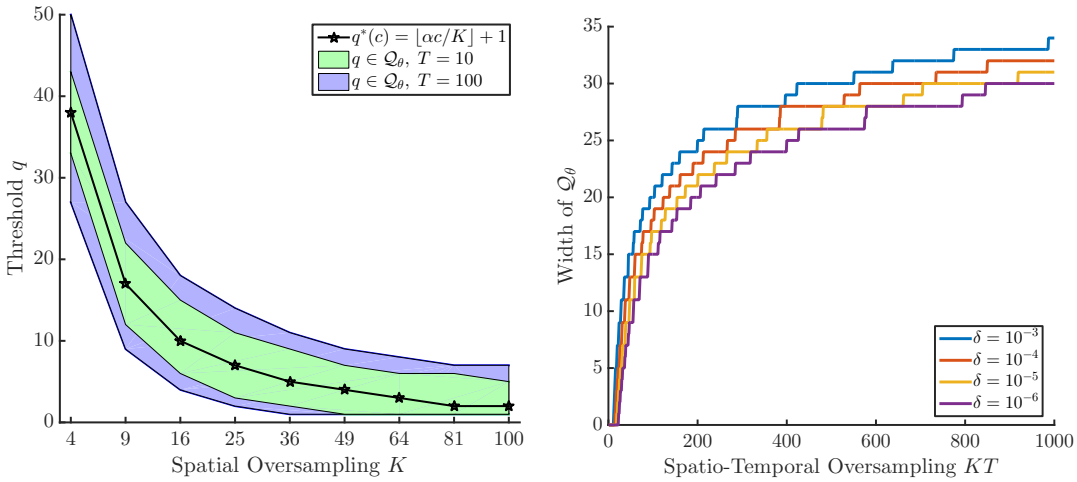
(a) $T = 10$, SNR $\in [28.93, 29.57]$ (b) $T = 25$, SNR $\in [32.98, 33.74]$

Fig. A.5. Phase transition for $T = 10$ and $T = 25$. SNR range is shown for average bit density $1 - \mathbb{E}[\gamma_q(c)]$ in the range $[0.264, 0.630]$. For all cases, we set $\delta = 2 \times 10^{-4}$, and $K = 4$.



(a) $T = 50$, SNR $\in [36.15, 36.80]$ (b) $T = 100$, SNR $\in [39.35, 39.82]$

Fig. A.6. Phase transition for $T = 50$ and $T = 100$. SNR range is shown for average bit density $1 - \mathbb{E}[\gamma_q(c)]$ in the range $[0.264, 0.630]$. For all cases, we set $\delta = 2 \times 10^{-4}$, and $K = 4$.



(a) (b)

Fig. A.7. (a) The threshold q and $\mathcal{Q}_\theta$ as K increases. (b) The width of $\mathcal{Q}_\theta$ as KT and δ changes.

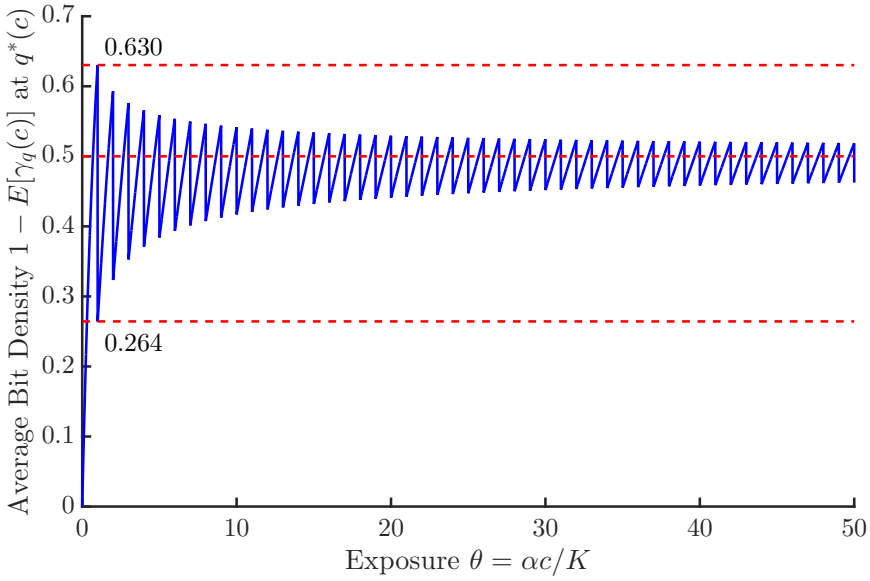


Fig. A.8. Average bit density $1 - \mathbb{E}[\gamma_q(c)]$ calculated at optimal threshold $q^* = \lfloor \theta \rfloor + 1$.

A.5 Influence of Non-Boxcar Kernel G

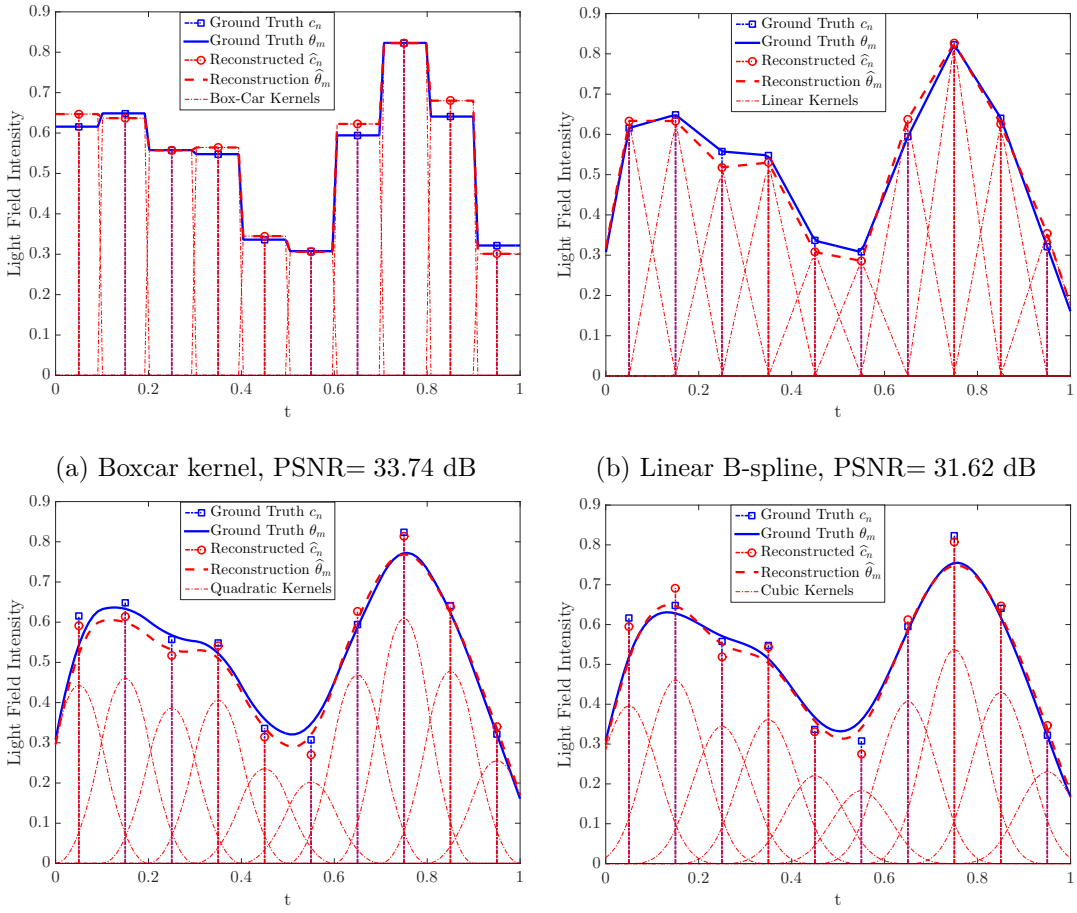
In this section, we discuss the boxcar kernel assumption in QIS model, i.e., $\boldsymbol{G} = \frac{1}{K}\boldsymbol{I}_{N\times N} \otimes \mathbf{1}_{K\times 1}$. We also study the effect of assuming a general kernel $\boldsymbol{G}$ on our results.

On QIS, we typically assume that there are micro-lenses on top of each jot or a group of jots. These micro-lenses ensure that the incident light converges onto the sensing site with no (or very minor) interference with adjacent jots or groups. As a result, we can model the incoming light using the boxcar kernel. This assumption is perhaps strong in some perspective, but it allows us to significantly simplify the theory and offer efficient implementations.

What if there is a mismatch between the physical model (e.g., using B-spline or Gaussian kernel $\boldsymbol{G}$) and the reconstruction (e.g., using boxcar)? To see the effect of this mismatch on the reconstruction quality, we conduct two sets of experiments.

- **1D Signal**: We consider a 1D signal with 10 coefficients. These 10 coefficients are modulated with boxcar kernels and B-spline kernels to generate two sets of incident light. On the QIS simulator, we set the spatial and temporal over-sampling factors as $K = 9$ and $T = 30$, respectively. Then we use the oracle threshold map for quantization. To reconstruct the images, we use boxcar kernel for both cases so that we have one matching case and one mismatching case. Figure A.9 shows the reconstructed signals. As expected, when the forward model matches with the reconstruction model, the reconstructed image has the highest PSNR. However, the gap between the cases are not significant.
- **2D Signal**: Figure A.10 shows a 2D example. Similar to the 1D case, boxcar kernel leads to the best reconstruction but its gap with the other cases are not significant.

The reader might think why we do not use B-spline on the reconstruction so that it will match with the forward model? In principle this is doable, but we need an iterative algorithm to compute the ML estimate such as gradient descent as reported in [6]. In contrast, the boxcar assumption allows us to use a closed-form ML estimate, which is practically much more affordable.



(a) Boxcar kernel, PSNR= 33.74 dB

(b) Linear B-spline, PSNR= 31.62 dB

(c) Quadratic B-spline, PSNR= 31.66 dB

(d) Cubic B-spline, PSNR= 32.15 dB

Fig. A.9. Spatial oversampling $K = 9$. Temporal oversampling $T = 30$. Oracle threshold map is used for quantization. Different kernels are used in synthesis and boxcar kernel is used in reconstruction. ML closed-form is used for reconstruction

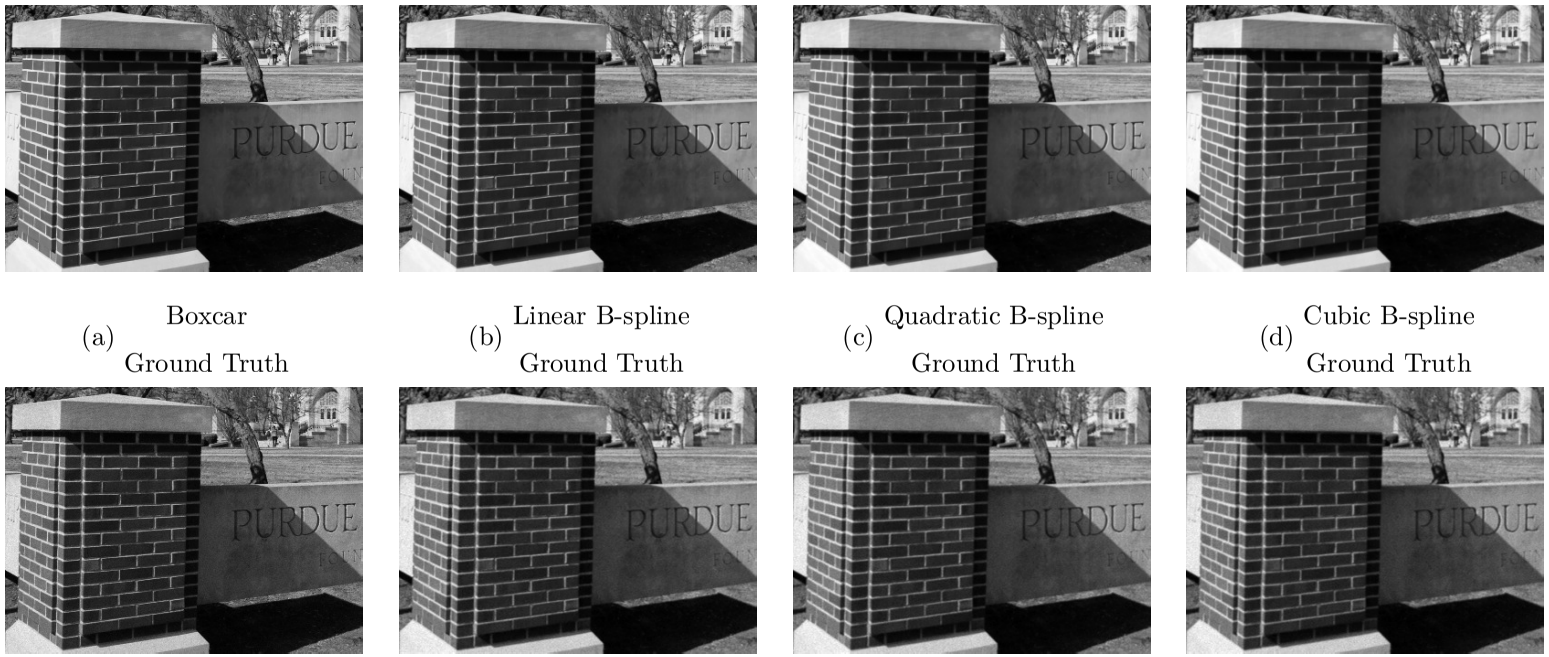


Fig. A.10. Ground truth and reconstructed images using simulated binary measurements synthesized by (a)(e) Boxcar, (b)(f) Linear B-spline, (c)(g) Quadratic B-spline, and (d)(h) Cubic B-spline kernels. In this experiment, we spatially oversample each pixel by $K = 4 \times 4$ binary bits and we use $T = 15$ independent temporal measurements. We use 8 frames for learning the threshold map using bisection method, and the remaining 7 frames are used for image reconstruction using the ML closed-form by the boxcar kernel assumption.

A.6 Supplementary HDR results

In this section, we show more results for HDR image reconstruction using our method compared to the fixed threshold approach. Figure A.11 show reconstructed HDR images using adapted Q-map by the bisection algorithm, and fixed Q-maps with low threshold ($q = 1$) and high threshold ($q_{\max} = 16$). The spatial and temporal oversampling factors are $K = 4$, and $T = 13$, respectively. Sensor gain is $\alpha = K^2/(q_{\max} - 1)$.

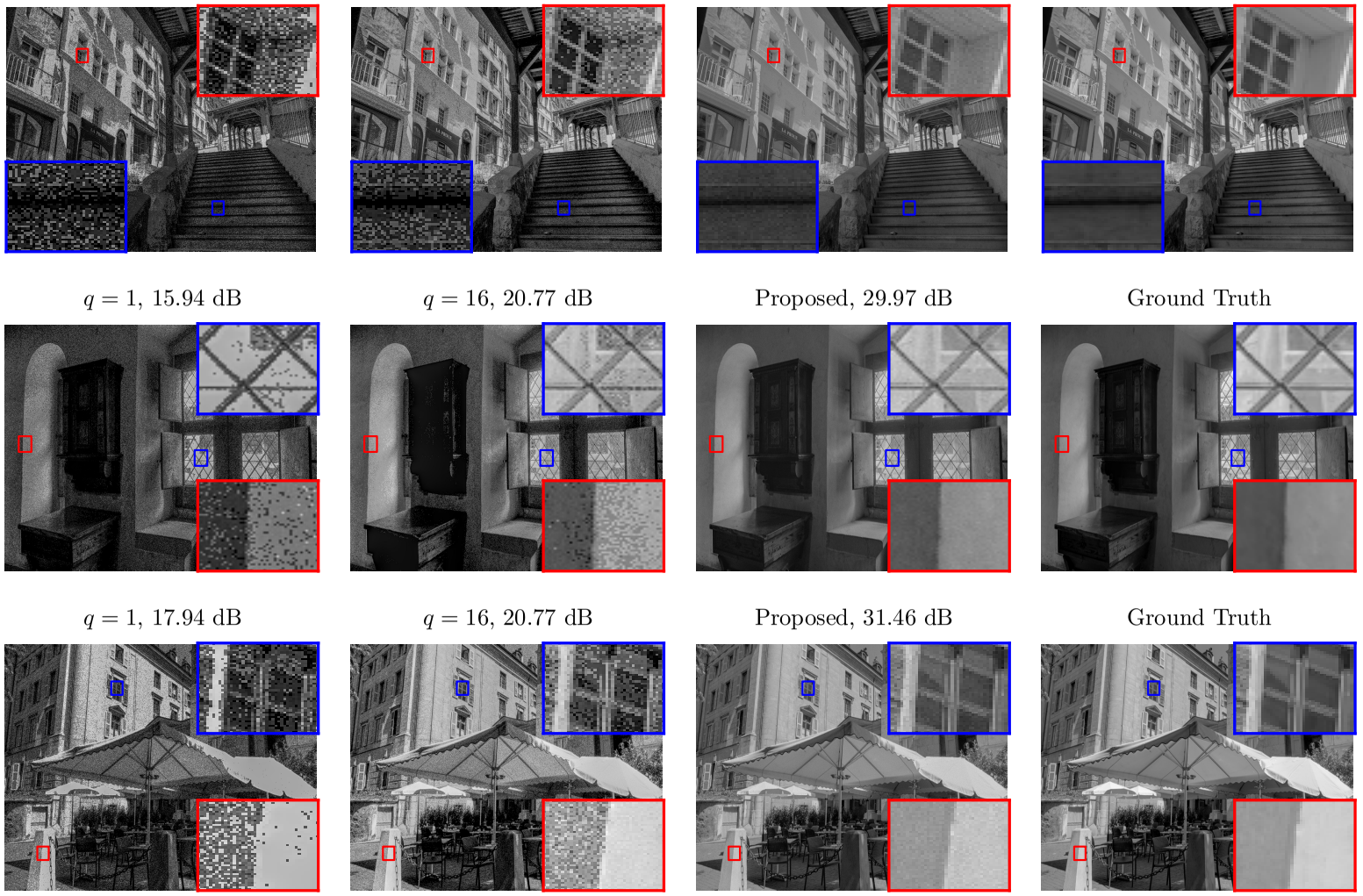


Fig. A.11. Reconstructed HDR images using different threshold maps

B. SUPPLEMENTARY MATERIAL FOR CHAPTER 5

This supplementary report provides the following additional information for Chapter 5

- Luminance/Chrominance Transformation Matrices of Other CFAs
- An Iterative Demosaicking Algorithm using ADMM
- Color Image Reconstruction using ADMM
- Color-Noise Trade-off

B.1 Luminance/Chrominance Transformation Matrices of Other CFAs

Algorithm II in the main manuscript performs demosaicking by frequency selection with the assumption of orthogonality. However, the CFAs proposed in [86], [93] and [98] do not satisfy the orthogonality constraint with our choice of $\boldsymbol{T}$ [95]. In this section, we derive for every CFA the transformation matrix $\boldsymbol{T}$ that makes its luminance and chrominance channel orthogonal so that we can apply Algorithm II.

Following the symbolic DFT method in [86], the frequency structure of RGBCY CFA proposed in [93] has the following form:

$$\frac{1}{16}\begin{bmatrix} 3B+10G+3R & 2R-2B & B-2G+R & 2R-2B \\ 2R-2B & B-2G+R & 0 & B-2G+R \\ B-2G+R & 0 & 2G-B-R & 0 \\ 2R-2B & B-2G+R & 0 & B-2G+R \end{bmatrix} \tag{B.1}$$

Hence, we can choose the luminance/chrominance transformation as

$$\begin{bmatrix} L \\ \alpha \\ \beta \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 3 & 10 & 3 \\ 1 & -2 & 1 \\ 2 & 0 & -2 \end{bmatrix}\begin{bmatrix} R \\ G \\ B \end{bmatrix} \leftrightarrow \boldsymbol{T}_{\text{RGBCY}} = \frac{1}{16}\begin{bmatrix} 3 & 10 & 3 \\ 1 & -2 & 1 \\ 2 & 0 & -2 \end{bmatrix} \tag{B.2}$$

As a result, the frequency structure is orthogonal where every chrominance component is modulated on distinct carrier as shown in Figure B.1 and shown in the following matrix representation

$$\frac{1}{16}\begin{bmatrix} L & \beta & \alpha & \beta \\ \beta & \alpha & 0 & \alpha \\ \alpha & 0 & -\alpha & 0 \\ \beta & \alpha & 0 & \alpha \end{bmatrix} \tag{B.3}$$

To ensure fairness between different CFAs, we normalize the matrix rows to unity so that all luminance and chrominance have the same noise power. To this end, the transformation matrix of RGBCY CFA can be written as

$$\boldsymbol{T}_{\text{RGBCY}} = \begin{bmatrix} \frac{3}{\sqrt{118}} & \frac{10}{\sqrt{118}} & \frac{3}{\sqrt{118}} \\ \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \end{bmatrix} \tag{B.4}$$

Similarly, we can do the same steps for RGBCWY CFA in [93] to obtain the following transformation matrix.

$$\boldsymbol{T}_{\text{RGBCWY}} = \begin{bmatrix} \frac{13}{\sqrt{822}} & \frac{22}{\sqrt{822}} & \frac{13}{\sqrt{822}} \\ \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \end{bmatrix} \tag{B.5}$$

As for Bayer CFA, and the CFA in [98], we use the following transformation matrix

$$\boldsymbol{T}_{\text{Bayer}} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \end{bmatrix} \tag{B.6}$$

Finally, for the CFA in [86], we use the following transformation matrix

$$\boldsymbol{T} = \begin{bmatrix} \frac{2}{\sqrt{22}} & \frac{3}{\sqrt{22}} & \frac{3}{\sqrt{22}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} \tag{B.7}$$

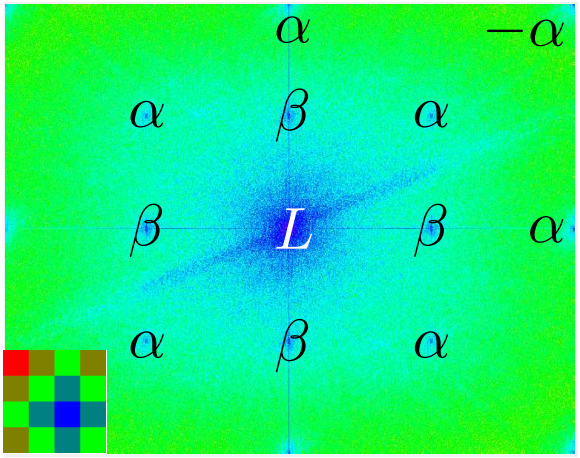


Fig. B.1. Frequency structure of RGBCY CFA [93] using the luminance/chrominance transformation (B.3)

B.2 Iterative Demosaicking Algorithm using ADMM

In the main manuscript, we modeled the color filter sampling by the following equation:

$$\begin{aligned}\theta(m,n) &= \alpha \boldsymbol{c}_{rgb}(m,n)^T \mathbf{im}_{rgb}(m,n) \\ &= \alpha \sum_{i\in\{r,g,b\}} c_i(m,n)\,\mathrm{im}_i(m,n).\end{aligned} \tag{B.8}$$

To write this equation in matrix form, we stack the vectorized color channels of the latent image im_i in long $3HW \times 1$ vector, and the mosaic channel $\boldsymbol{\theta}$ in long $HW \times 1$ vector as follows:

$$\mathbf{im} \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{vec}(\mathrm{im}_r) \\ \mathrm{vec}(\mathrm{im}_g) \\ \mathrm{vec}(\mathrm{im}_b) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\theta} \stackrel{\text{def}}{=} \mathrm{vec}(\theta). \tag{B.9}$$

We define the sampling matrix $\boldsymbol{C} \stackrel{\text{def}}{=} [\boldsymbol{C}_r, \boldsymbol{C}_g, \boldsymbol{C}_b] \in [0,1]^{HW\times 3HW}$, where $\boldsymbol{C}_i \stackrel{\text{def}}{=} \mathrm{diag}(\mathrm{vec}(c_i)), \forall i \in \{r,g,b\}$. Then, the color filter sampling can be written as:

$$\boldsymbol{\theta} = \alpha \boldsymbol{C}\,\mathbf{im}. \tag{B.10}$$

By recalling the forward model (B.10), we can write the inverse problem for obtaining the latent color image $\mathbf{im}$ from the light exposure on QIS $\boldsymbol{\theta}$ as follows

$$\widehat{\mathbf{im}} = \arg\min_{\mathbf{im}} ||\alpha \boldsymbol{C}\,\mathbf{im} - \boldsymbol{\theta}||_2^2 + \lambda g(\mathbf{im}), \tag{B.11}$$

where $\boldsymbol{C}$ is the color filter sampling operator. The first term in the cost function is a data-fidelity term that forces $\widehat{\mathbf{im}}$ to agree with the measurements $\boldsymbol{y}$. The second term is a regularization term to improve the conditioning of our ill-posed problem. λ is a positive scalar that controls the amount of regularization.

To solve the inverse problem (B.11), we may use any optimization toolbox since it is convex. Here, we report our results using the Plug-and-Play (PnP) ADMM algorithm [110], which has demonstrated effectiveness in image restoration tasks. Starting from an initial guess $\mathbf{im}^{(0)}$, the PnP ADMM algorithm iteratively updates its estimate via two steps:

Demosaicking Module:

$$\mathbf{im}^{(k+1)} = (\alpha^2 \boldsymbol{C}^T\boldsymbol{C} + \rho \boldsymbol{I})^{-1}(\alpha \boldsymbol{C}^T\boldsymbol{y} + \rho(\boldsymbol{v}^{(k)} - \boldsymbol{u}^{(k)})), \tag{B.12}$$

Denoising Module:

$$\boldsymbol{v}^{(k+1)} = \mathcal{D}_{\lambda/\rho}(\mathbf{im}^{(k+1)} + \boldsymbol{u}^{(k)}), \tag{B.13}$$

and updates the Lagrange multiplier by $\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} - (\mathbf{im}^{(k+1)} - \boldsymbol{v}^{(k+1)})$. For additional details on PnP ADMM, we refer the readers to, e.g., [110]. Here, ρ is an internal parameter that controls the convergence. The operator $\mathcal{D}$ is an off-the-shelf image denoiser, e.g., BM3D in our experiments. The subscript λ/ρ denotes the denoising strength, i.e., the hypothesized "noise variance". The inversion in the demosaicking module is performed in closed form because $\boldsymbol{C}^T\boldsymbol{C}$ exhibits a block diagonal structure.

The optimization problem in (B.11) does not take into account of the crosstalk effect. [1] Like most of the mainstream image and signal processing (ISP) pipelines, we reduce the cross-talk via a color correction step.

[1] In principle we can incorporate the crosstalk kernel into the $\boldsymbol{C}$ matrix but then $\boldsymbol{C}$ will have a complicated structure which does not allow simple inversion.

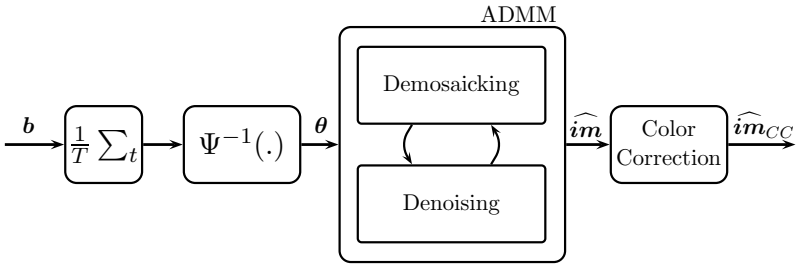


Fig. B.2. Block diagram of our reconstruction method. Given QIS binary frames $\boldsymbol{b}$, we obtain an approximately clean estimate for QIS light exposure $\boldsymbol{\theta}$. Afterwards, we apply an iterative ADMM algorithm for demosaicking. Finally, we do color correction to remove the crosstalk effect.

B.3 Color Image Reconstruction using ADMM

In this experiment, we perform color image reconstruction using the 24 and 18 color images in Kodak and McMaster datasets, respectively. QIS parameters are $q = 1$, $\alpha = 2$, and $T = 1000$. Color filtering is obtained using the proposed CFAs and other arrays proposed in literature [83,86,87,95,98]. For every CFA, we generate mosaicked images under two scenarios: 1) crosstalk kernels with leakage factors $(\alpha_r, \alpha_g, \alpha_b) = (0, 0, 0)$, i.e., no crosstalk, and 2) crosstalk kernels with leakage factors $(\alpha_r, \alpha_g, \alpha_b) = (0.23, 0.15, 0.10)$. Color correction is performed for the second scenario to remove crosstalk color de-saturation effect. For both scenarios, we apply 300 iterations of the Plug-and-Play ADMM algorithm for image demosaicking with BM3D denoising prior and $\rho = 1$.

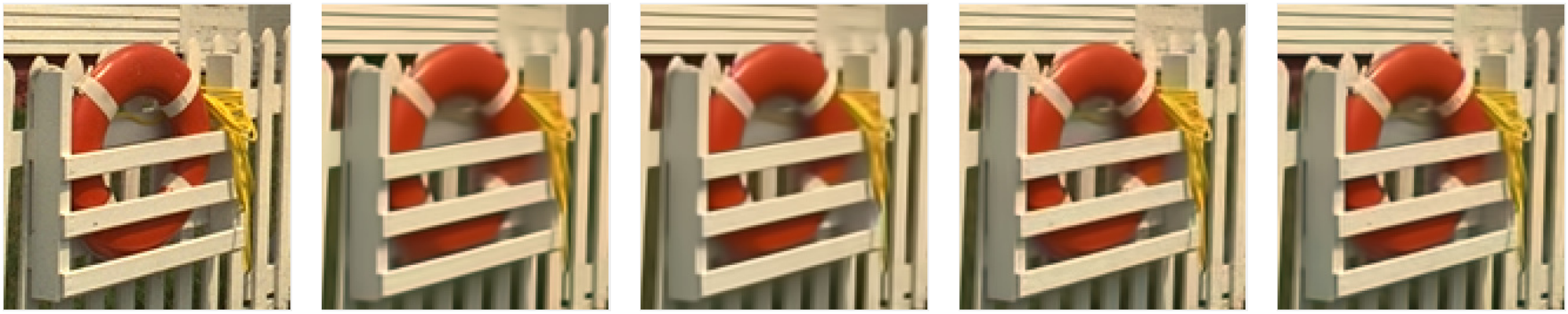
Different CFAs have different convergence properties according to the condition number of their corresponding masking matrix $\boldsymbol{C}$. Therefore, we perform fine-tuning for the λ parameter for every CFA and every color image. Specifically, we run the ADMM algorithm for 50 iterations using $\lambda \in \{0.005, 0.01, 0.015, 0.02, 0.025, 0.03\}$ and pick the λ that obtains the best color-PSNR. For McMaster dataset, we do the same fine-tuning, except that we run the ADMM algorithm for 100 iterations.

The last four columns in Table B.1 show the median PSNR of the 24 and 18 color images in Kodak and McMaster datasets, respectively. The scenarios of crosstalk and

Table B.1.
Reconstruction quality measured by median PSNR on Kodak and McMaster color datasets.

Size	CFA Pattern	CPSNR-McM		CPSNR-Kodak	
		w/o Ctk	w/ Ctk	w/o Ctk	w/ Ctk
4×4	Hao *et* al. [86]	21.69	26.81	27.92	29.68
	RGBCWY [93]	**30.07**	29.86	31.14	**30.80**
	Ours	29.94	**29.90**	**31.25**	30.45
3×3	Cheng *et* al. [98]	29.39	29.11	29.50	28.52
	Ours	**30.78**	**30.13**	**31.32**	**31.00**
3×2	Condat [95]	**31.13**	30.57	**33.29**	32.59
	Ours	28.37	**32.03**	33.22	**32.68**
4×2	Hirakawa-Wolfe [83]	26.49	30.23	31.59	31.28
	Ours	**26.72**	**30.70**	**32.04**	**32.01**

no crosstalk are denoted in the table as "w/ Ctk" and "w/o Ctk", respectively. We notice that our proposed CFAs achieves higher PSNR compared to other CFAs when crosstalk exists. This is attributed to their improved robustness to crosstalk compared to other arrays. Figure B.3 shows crops of reconstructed images using different CFAs. Images that are captured using our proposed CFAs show good amount of details, and good color fidelity.



Ground Truth | 4×4: [86], 31.07dB | 3×3: [98], 28.24dB | 3×2: [95], 33.3.dB | 4×2: [83], 32.47dB

4×4: [93], 31.15dB | 4×4: Ours, 32.13dB | 3×3: Ours, 32.39dB | 3×2: Ours, 33.93dB | 4×2: Ours, 32.60dB

Fig. B.3. Reconstructed color images from the QIS measurements. Each subfigure shows the result using a particular color filter array design. The reconstruction is based on the same ADMM algorithm with optimized parameters for each case. Thus, the PSNRs are the maximum-achievable values within the framework.

B.4 Color-Noise Trade-off

In this experiment, we compare the trade-off between noise amplification and color accuracy of our proposed CFAs and other CFAs in literature. To do so, we use the Macbeth color chart that comprises 24 color patches. The forward model consists of illumination using D65 light and mosaicking using a CFA and crosstalk using the crosstalk kernels:

$$g_i = \begin{bmatrix} 0 & \alpha_i/4 & 0 \\ \alpha_i/4 & 1-\alpha_i & \alpha_i/4 \\ 0 & \alpha_i/4 & 0 \end{bmatrix}, \; i \in \{r, g, b\}, \tag{B.14}$$

with $(\alpha_r, \alpha_g, \alpha_b) = (0.45, 0.30, 0.20)$ as suggested in [93]. QIS parameters are $q = 1$, $\alpha = 2$ and $T = 1000$. We use Algorithm II for demosaicking with frequency selection. The low pass filter is $m \times m$ Gaussian having standard deviation $\sigma = m/3$ and multiplied by a Hamming window to eliminate windowing effect. Since the ground truth color values of Macbeth color chart are known, we compute the color correction matrix $\boldsymbol{M}$ by solving the following regularized linear least squares optimization problem with white balance constraint:

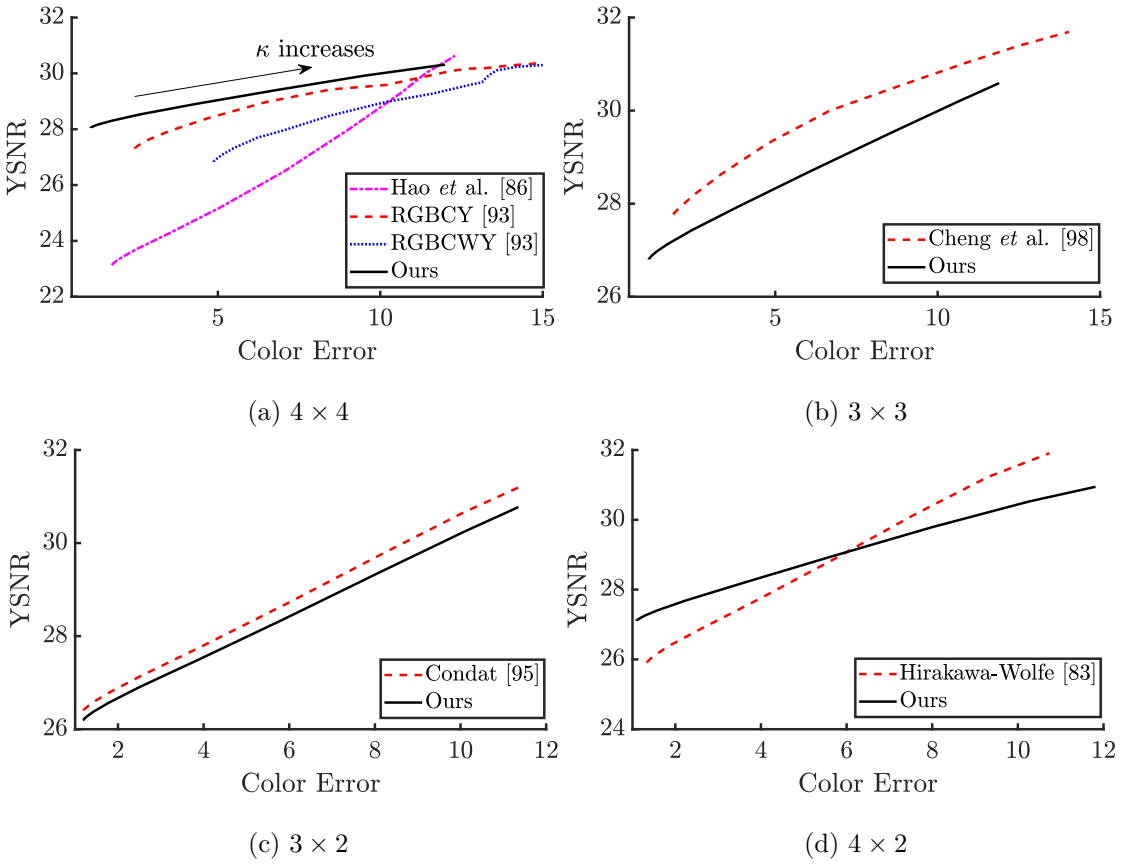
$$\begin{aligned} \boldsymbol{M} &= \arg\min_{\boldsymbol{M}} \epsilon_c(\boldsymbol{M}) + \kappa \sum_{i=1}^{24} ||\mathrm{Cov}(\boldsymbol{M}\boldsymbol{Q}^{(i)}_{\mathrm{False}})||_2^2 \\ &\text{subject to} \\ \boldsymbol{M}\boldsymbol{u} &= \boldsymbol{u} \end{aligned} \tag{B.15}$$

where $\epsilon_c(\boldsymbol{M}) = \mathrm{Tr}\left\{ (\boldsymbol{M}\boldsymbol{Q}_{\mathrm{False}} - \boldsymbol{Q}_{\mathrm{GT}})^T (\boldsymbol{M}\boldsymbol{Q}_{\mathrm{False}} - \boldsymbol{Q}_{\mathrm{GT}}) \right\}$ is the color error. $\boldsymbol{Q}_{\mathrm{False}}$ and $\boldsymbol{Q}_{\mathrm{GT}}$ are $3 \times K$ matrices containing the measured color values and the corresponding ground truth color values of K pixels. $\boldsymbol{u} \stackrel{\mathrm{def}}{=} [0.95, 1, 1.0889]^T$ is the white point for D65 illuminant.

To draw the noise-color trade-off curve, we vary the parameter κ in (B.15) from 0 to 10^{10} on the log-scale. Color error is quantified with the CIEDE2000 metric which is obtained by calculating the mean square color difference in the CIELAB color space [93]. Visual noise is measured by the YSNR metric as defined in ISO

12232 [93]. To ensure that we obtain the best possible performance of every CFA and κ, we repeat Algorithm II with different sizes of the low pass filter $m \in \{15, 17, \ldots, 25\}$ and pick the value that maximizes YSNR and minimizes color error. Since YSNR should be increased and color error should be decreased, the tradeoff curve is better when it is shifted to upper left.

Figure B.4 shows the trade-off curves for the proposed CFAs and other CFAs. Our 4×4 CFA is better than other 4×4 CFAs for almost all values of κ. Our 3×3 CFA achieves lower color error compared to [98]. As for 4×2 CFAs, our CFA is better than [83] if we restrict to small color error. However, if we allow larger color error, then [83] is better. For the 3×2 case, Condat CFA [95] is better than hours for values of $kappa > 0$, but our CFA achieves better performance on natural images as mentioned in Experiment 3 in the main manuscript.



(a) 4×4 (b) 3×3

(c) 3×2 (d) 4×2

Fig. B.4. Color-Noise trade-off for different CFAs. Demosaicking is performed using Algorithm II. κ in (B.15) is varied from 0 to 10^{10}.

C. PROOFS

C.1 Proof of Proposition 3.1.2

By using the partitioning in (3.12) and substituting with the constraint from (3.13), we can decompose (3.1) into a triple sum formula:

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{t=0}^{T-1} \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} \left\{ b_{Kn+k,t} \log \left(1 - \Psi_q \left(\frac{\alpha c_n}{K} \right) \right) + (1 - b_{Kn+k,t}) \log \Psi_q \left(\frac{\alpha c_n}{K} \right) \right\}, \tag{C.1}$$

Let $\mathcal{B}_{n,t} \stackrel{\text{def}}{=} \{b_{Kn,t}, \ldots, b_{Kn+(K-1),t}\}$ be defined as the n-th block of the t-th frame, and

$$S_n \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} b_{Kn+k,t} \tag{C.2}$$

be defined as the sum of the bits (i.e., the number of one's) in $\mathcal{B}_{n,t}$. Then, (C.1) becomes

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{n=0}^{N-1} S_n \log \left(1 - \Psi_q \left(\frac{\alpha c_n}{K} \right) \right) + (L - S_n) \log \Psi_q \left(\frac{\alpha c_n}{K} \right), \tag{C.3}$$

where $L = KT$. By observing (C.3), we notice that it can be decomposed into N subproblems as follows.

$$\widehat{c}_n = \underset{c_n}{\operatorname{argmax}} \; S_n \log \left(1 - \Psi_q \left(\frac{\alpha c_n}{K} \right) \right) + (L - S_n) \log \Psi_q \left(\frac{\alpha c_n}{K} \right), \tag{C.4}$$

where $n \in \{0, \ldots, N-1\}$. By setting the first derivative to zero, we can obtain $\widehat{c}_n$ as follows.

$$\left(\frac{S_n}{1 - \Psi_q \left(\frac{\alpha c_n}{K} \right)} - \frac{L - S_n}{\Psi_q \left(\frac{\alpha c_n}{K} \right)} \right) \left(-\frac{\alpha}{K} \frac{e^{-\frac{\alpha c_n}{K}} \left(\frac{\alpha c_n}{K} \right)^{q-1}}{\Gamma(q)} \right) = 0.$$

Since the second bracket is non-zero, we can divide both sides by it and rearrange the terms to get

$$\Psi_q \left(\frac{\alpha c_n}{K} \right) = 1 - \frac{S_n}{L}.$$

which can be solved in c_n using the inverse incomplete Gamma function to give:

$$\widehat{c}_n = \frac{K}{\alpha}\Psi_q^{-1}\left(1 - \frac{S_n}{L}\right), \qquad n = 0, \ldots, N-1, \tag{C.5}$$

C.2 Proof of Theorem 3.3.1

For notational simplicity we drop the subscript n. Our goal is to show that if $X \sim \text{Binomial}(L, p)$, then the transformed variable

$$\mathcal{T}(X) = \sqrt{L + \frac{1}{2}}\sin^{-1}\left(\sqrt{\frac{X + \frac{3}{8}}{L + \frac{3}{4}}}\right) \tag{C.6}$$

has a variance $\text{Var}[\mathcal{T}(X)] = \frac{1}{4} + \mathcal{O}(L^{-2})$. To this end, we first consider the function $\mathcal{Q}$ such that

$$\mathcal{Q}(X) = \mathcal{T}(X) - \sqrt{L + \frac{1}{2}}\sin^{-1}\sqrt{p}.$$

Since $\text{Var}[\mathcal{Q}(X) + c] = \text{Var}[\mathcal{Q}(X)]$ for any c, by letting $c = -\sqrt{L + \frac{1}{2}}\sin^{-1}\sqrt{p}$ we observe that showing Theorem 2 is equivalent to showing $\text{Var}[\mathcal{Q}(X)] = \frac{1}{4} + \mathcal{O}(L^{-2})$.

To show the desired result, we note that for any α and β, the arcsin function has the property that

$$\sin^{-1}\alpha - \sin^{-1}\beta = \sin^{-1}\left(\alpha\sqrt{1-\beta^2} - \beta\sqrt{1-\alpha^2}\right).$$

Define $F \stackrel{\text{def}}{=} \frac{X+\frac{3}{8}}{L+\frac{3}{4}}$, and substitute $\alpha = \sqrt{F}$, $\beta = \sqrt{p}$, it follows that

$$\mathcal{Q}(X) = \sqrt{L + \frac{1}{2}}\sin^{-1}\left(\sqrt{(1-p)F} - \sqrt{p(1-F)}\right). \tag{C.7}$$

There are two terms in this equation. The first term $\sqrt{L + \frac{1}{2}}$ can be expanded (using Taylor expansion) to its first second order as

$$\sqrt{L + \frac{1}{2}} = \sqrt{L}\left(1 + \frac{1}{2L}\right)^{\frac{1}{2}} = \sqrt{L}\left(1 + \frac{1}{4L} + \mathcal{O}(L^{-2})\right).$$

The arcsin function can be expanded to its second order as

$$\sin^{-1} W = W + \frac{W^3}{6} + \frac{3W^5}{40} + \frac{5W^7}{112} + \ldots,$$

for $W = \sqrt{(1-p)F} - \sqrt{p(1-F)}$.

We next consider the standardized binomial random variable by defining

$$Y \stackrel{\text{def}}{=} \frac{X - Lp}{\sqrt{Lp(1-p)}}. \tag{C.8}$$

Then, by Lemma 1, it follows that

$$\begin{aligned} W &= \sqrt{(1-p)F} - \sqrt{p(1-F)} \\ &= \frac{Y}{2\sqrt{L}} + \frac{(2p-1)(2Y^2-3)}{16L\sqrt{p(1-p)}} + \frac{-16Y^3p^2 + 16Y^3p - 6Y^3 + 9Y}{96L^{\frac{3}{2}}p(1-p)} + \mathcal{O}(L^{-2}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{Q}(X) &= \sqrt{L}\left(1 + \frac{1}{4L} + \mathcal{O}(L^{-2})\right)\left(W + \frac{W^3}{6} + \mathcal{O}(W^5)\right) \\ &= a_0 + a_1 Y + a_2 Y^2 + a_3 Y^3 + \mathcal{O}(Y^5), \end{aligned}$$

where

$$\begin{aligned} a_0 &= -\tfrac{3(2p-1)}{16\sqrt{Lp(1-p)}}, \quad a_1 = \tfrac{1}{2} + \tfrac{1}{8L} - \tfrac{3}{32Lp(1-p)} \\ a_2 &= \tfrac{2p-1}{8\sqrt{Lp(1-p)}}, \qquad a_3 = \tfrac{16p^2-16p+6}{96Lp(1-p)}. \end{aligned}$$

Since the first four moments of Y are

$$\mathbb{E}[Y] = 0, \;\; \mathbb{E}[Y^2] = 1, \;\; \mathbb{E}[Y^3] = -\frac{2p-1}{\sqrt{Lp(1-p)}}, \;\; \mathbb{E}[Y^4] = 3 + \frac{1 - 6p(1-p)}{Lp(1-p)},$$

we conclude that

$$\begin{aligned} \mathrm{Var}[\mathcal{Q}(X)] &= a_1^2\mathrm{Var}[Y] + a_2^2\mathrm{Var}[Y^2] + 2a_1a_2\mathrm{Var}[Y^3] + 2a_1a_3\mathrm{Var}(Y^4) \\ &= a_1^2 - a_2^2 + 2a_1a_2\mathbb{E}[Y^3] + (2a_1a_3 + a_2^2)\mathbb{E}[Y^4] = \frac{1}{4} + \mathcal{O}(L^{-2}). \end{aligned}$$

Lemma 1 *Let* $F = \frac{X+\frac{3}{8}}{L+\frac{3}{4}}$ *and* $Y = \frac{X-Lp}{\sqrt{Lp(1-p)}}$. *It holds that*

$$\begin{aligned} \sqrt{(1-p)F} = \sqrt{p(1-p)} &+ \frac{(1-p)Y}{2\sqrt{L}} - \frac{\sqrt{1-p}(6p + 2(1-p)Y^2 - 3)}{16L\sqrt{p}} \\ &- \frac{(1-p)Y(6p - 2(1-p)Y^2 + 3)}{32pL^{\frac{3}{2}}} + \mathcal{O}(L^{-2}). \end{aligned} \tag{C.9}$$

$$\begin{aligned} \sqrt{p(1-F)} = \sqrt{p(1-p)} &- \frac{pY}{2\sqrt{L}} - \frac{\sqrt{p}(-6p + 2pY^2 + 3)}{16L\sqrt{1-p}} \\ &- \frac{pY(6p + 2pY^2 - 9)}{32(1-p)L^{\frac{3}{2}}} + \mathcal{O}(L^{-2}). \end{aligned} \tag{C.10}$$

Proof Note that $Y = \frac{X-Lp}{\sqrt{Lp(1-p)}}$ is equivalent to $X = Y\sqrt{Lp(1-p)} + Lp$. Thus, F can be expressed in terms of Y as

$$F = \frac{\left(Y\sqrt{Lp(1-p)} + Lp\right) + \frac{3}{8}}{L + \frac{3}{4}} = \left(Y\sqrt{\frac{p(1-p)}{L} + p + \frac{3}{8L}}\right)\left(1 + \frac{3}{4L}\right)^{-1}.$$

For large L, we have $\frac{3}{4L} \ll 1$. Thus, by expanding $\left(1 + \frac{3}{4L}\right)^{-1}$ we have

$$F = \left(Y\sqrt{\frac{p(1-p)}{L} + p + \frac{3}{8L}}\right)\left(1 - \frac{3}{4L} + \mathcal{O}(L^{-2})\right) = p\left(1 + E_1\right),$$

where

$$E_1 = \sqrt{\frac{1-p}{p}}\frac{Y}{\sqrt{L}} - \frac{\frac{3}{4} - \frac{3}{8p}}{L} - \sqrt{\frac{p}{1-p}}\frac{3Y}{4L^{\frac{3}{2}}} + \mathcal{O}(L^{-2}).$$

By expanding $\sqrt{1+E_1}$, we arrive at

$$\sqrt{F} = \sqrt{p}\sqrt{1+E_1} = \sqrt{p}\left(1 + \frac{E_1}{2} - \frac{E_1^2}{8} + \frac{E_1^3}{16} + \mathcal{O}(E_1^4)\right).$$

Multiplying both sides by $\sqrt{1-p}$ and substituting for E_1 yields

$$\begin{aligned}\sqrt{(1-p)F} = \sqrt{p(1-p)} + \frac{(1-p)Y}{2\sqrt{L}} - \frac{\sqrt{1-p}(6p + 2(1-p)Y^2 - 3)}{16L\sqrt{p}} \\ - \frac{(1-p)Y(6p - 2(1-p)Y^2 + 3)}{32pL^{\frac{3}{2}}} + \mathcal{O}(L^{-2}).\end{aligned}$$

The proof of the second equality can be done by expressing $1 - F$ in terms of Y as

$$\begin{aligned}1 - F &= \left(-Y\sqrt{\frac{p(1-p)}{L}} + (1-p) + \frac{3}{8L}\right)\left(1 + \frac{3}{4L}\right)^{-1}. \\ &= \left(-Y\sqrt{\frac{p(1-p)}{L}} + (1-p) + \frac{3}{8L}\right)\left(1 - \frac{3}{4L} + \mathcal{O}(L^{-2})\right) = (1-p)\left(1 + E_2\right),\end{aligned}$$

where

$$E_2 = -\sqrt{\frac{p}{1-p}}\frac{Y}{\sqrt{L}} - \frac{\frac{3}{4} - \frac{3}{8(1-p)}}{L} + \sqrt{\frac{1-p}{p}}\frac{3Y}{4L^{\frac{3}{2}}} + \mathcal{O}(L^{-2}).$$

By expanding $\sqrt{1+E_2}$, we arrive at

$$\sqrt{1-F} = \sqrt{1-p}\sqrt{1+E_2} = \sqrt{1-p}\left(1 + \frac{E_2}{2} - \frac{E_2^2}{8} + \frac{E_2^3}{16} + \mathcal{O}(E_2^4)\right).$$

Multiplying both sides by $\sqrt{p}$ and substituting for E_2 yields

$$\sqrt{p(1-F)} = \sqrt{p(1-p)} - \frac{pY}{2\sqrt{L}} - \frac{\sqrt{p}(-6p+2pY^2+3)}{16L\sqrt{1-p}} - \frac{pY(6p+2pY^2-9)}{32(1-p)L^{\frac{3}{2}}} + \mathcal{O}(L^{-2}).$$

■

C.3 Proof of Proposition 4.1.2

The Fisher Information metric is defined as:

$$I_q(c) \stackrel{\text{def}}{=} \mathbb{E}_B\left[\frac{-\partial^2}{\partial c^2}\log \mathbb{P}(B=b;\theta,q)\right], \tag{C.11}$$

where $\theta = \alpha c/K$. Using the chain rule, we can derive the Fisher Information as follows

$$I_q(c) = \left(\frac{\alpha}{K}\right)^2 \mathbb{E}_B\left[\frac{-\partial^2}{\partial \theta^2}\log \mathbb{P}(B=b;\theta,q)\right]. \tag{C.12}$$

The expectation can be calculated as follows

$$\begin{aligned} I_q(c) = &\left(\frac{\alpha}{K}\right)^2 \left[\frac{-\partial^2}{\partial \theta^2}\log \mathbb{P}(B=1;\theta,q)\right] \mathbb{P}(B=1;\theta,q) \\ &+ \left(\frac{\alpha}{K}\right)^2 \left[\frac{-\partial^2}{\partial \theta^2}\log \mathbb{P}(B=0;\theta,q)\right] \mathbb{P}(B=0;\theta,q) \end{aligned} \tag{C.13}$$

Using (2.16) to differentiate the 1st term, we get:

$$\begin{aligned} &\frac{\partial^2}{\partial \theta^2}\log \mathbb{P}(B=1;\theta,q) = \frac{\partial^2}{\partial \theta^2}\log\left(1-\Psi_q(\theta)\right) \\ &= \frac{R'(1-\Psi_q(\theta)) - R^2/\Gamma(q)}{\Gamma(q)\left(1-\Psi_q(\theta)\right)^2}, \end{aligned} \tag{C.14}$$

where $R = e^{-\theta}\theta^{q-1}$ and $R' = \partial R/\partial\theta$. Similarly, the second term is

$$\begin{aligned} &\frac{\partial^2}{\partial \theta^2}\log \mathbb{P}(B=0;\theta,q) = \frac{\partial^2}{\partial \theta^2}\log \Psi_q(\theta) \\ &= -\frac{R'\Psi_q(\theta) + R^2/\Gamma(q)}{\Gamma(q)\left(\Psi_q(\theta)\right)^2}. \end{aligned} \tag{C.15}$$

Substitute (C.14) and (C.15) in (C.13) yields

$$\begin{aligned} I_q(\theta) &= \left(\frac{\alpha}{K}\right)^2 \Big[- \frac{R'\Gamma(q)(1-\Psi_q(\theta)) - R^2}{\Gamma^2(q)\,(1-\Psi_q(\theta))} \\ &\quad + \frac{R'\Gamma(q)\Psi_q(\theta) + R^2}{\Gamma^2(q)\Psi_q(\theta)} \Big] \\ &= \left(\frac{\alpha}{K}\right)^2 \frac{e^{-2\theta}\theta^{2q-2}}{\Gamma^2(q)\Psi_q(\theta)\,(1-\Psi_q(\theta))}. \end{aligned}$$

C.4 Proof of Proposition 4.1.3

The lower bound is obtained by observing that the product $\Psi_q(\theta)\,(1-\Psi_q(\theta))$ attains its maximum value when $\Psi_q(\theta) = 1/2$. Substituting with the upper bound $\Psi_q(\theta)\,(1-\Psi_q(\theta)) \leq 1/4$, we get:

$$\begin{aligned} \log(c^2 I_q(c)) &= \log\left\{ \left(\frac{\alpha c}{K}\right)^2 \frac{e^{-2\theta}\theta^{2q-2}}{\Gamma^2(q)\Psi_q(\theta)\,(1-\Psi_q(\theta))} \right\} \\ &= \log \frac{e^{-2\theta}\theta^{2q}}{\Gamma^2(q)\Psi_q(\theta)\,(1-\Psi_q(\theta))} \\ &\geq \log \frac{4e^{-2\theta}\theta^{2q}}{\Gamma^2(q)} \\ &= 2\log 2 - 2\theta + 2q\log\theta - 2\log\Gamma(q) \\ &= 2\left(\log 2 - \frac{\alpha c}{K} + q\log\frac{\alpha c}{K} - \log\Gamma(q)\right). \end{aligned}$$

C.5 Proof of Proposition 4.1.4

Using the definition of Gamma function $\Gamma(q) = (q-1)!$ and $\theta = \frac{\alpha c}{K}$, we can rewrite the lower bound in Proposition 4.1.3 as follows.

$$\begin{aligned} L_q(c) &= 2\left(\log 2 - \theta + q\log\theta - \log(q-1)!\right) \\ &= 2\left(\log 2 - \theta + (q-1)\log\theta + \log\theta - \log\prod_{k=1}^{q-1} k\right) \\ &= 2\left(\log 2 - \theta + \sum_{k=1}^{q-1}\log(\theta/k) + \log\theta\right) \end{aligned}$$

The only dependence on q is in the second term, so we take a closer look at it. When $q - 1 < \lfloor\theta\rfloor$, all summands $\log(\theta/k)$ are positive because $k < \lfloor\theta\rfloor$. Hence, the total

sum increases by increasing q. On the other hand, when $q-1 > \lfloor\theta\rfloor$, we start to add negative summands $\log(\theta/k)$ because $k > \theta$. Therefore, the total sum decreases on increasing $q-1$ over $\lfloor\theta\rfloor$. Thus, maximum is obtained at $q = \lfloor\theta\rfloor + 1 = \lfloor\frac{\alpha c}{K}\rfloor + 1$.

C.6 Proof of Proposition 4.2.1

By definition, $S \stackrel{\text{def}}{=} \sum_{t=0}^{T-1}\sum_{k=0}^{K-1} B_{k,t}$ is the summation of KT independent i.i.d. Bernoulli random variables. Therefore, S is a binomial random variable with parameters $n \stackrel{\text{def}}{=} KT$ and $p \stackrel{\text{def}}{=} 1-\Psi(\alpha c/K)$. The mean and variance of a binomial random variable is $\mathbb{E}[S] = np$, and $\text{Var}[S] = np(1-p)$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\gamma_q(c)\right] &= 1 - \frac{\mathbb{E}[S]}{KT} = \Psi_q\left(\frac{\alpha c}{K}\right), \quad \text{and} \\
\text{Var}\left[\gamma_q(c)\right] &= \frac{\text{Var}\left[S\right]}{K^2T^2} = \frac{1}{KT}\Psi_q\left(\frac{\alpha c}{K}\right)\left(1-\Psi_q\left(\frac{\alpha c}{K}\right)\right).
\end{aligned}$$

C.7 Proof of Proposition 4.2.2

The probability $\mathbb{P}[0 < \gamma_q(c) < 1]$ can be evaluated by checking the complement when $\gamma_q(c) = 0$ or $\gamma_q(c) = 1$:

$$\begin{aligned}
\mathbb{P}[0 < \gamma_q(c) < 1] &= 1 - \mathbb{P}[\gamma_q(c) = 0] - \mathbb{P}[\gamma_q(c) = 1] \\
&= 1 - \mathbb{P}[S = 0] - \mathbb{P}[S = KT] \\
&\stackrel{(a)}{=} 1 - \Psi_q(\theta)^{KT} - [1-\Psi_q(\theta)]^{KT},
\end{aligned}$$

where (a) follows from the fact that S, which is a sum of i.i.d. Bernoulli random variables, is a binomial random variable.

Let $0 < \delta < 1$. If

$$1 - \left(\frac{\delta}{2}\right)^{\frac{1}{KT}} \leq \Psi_q(\theta) \leq \left(\frac{\delta}{2}\right)^{\frac{1}{KT}},$$

then we have

$$\Psi_q(\theta)^{KT} < \frac{\delta}{2} \text{ and } [1-\Psi_q(\theta)]^{KT} < \frac{\delta}{2}.$$

Thus, it holds that

$$1 - \Psi_q(\theta)^{KT} - [1 - \Psi_q(\theta)]^{KT} > 1 - \delta.$$

C.8 Proof of Proposition 5.2.1

Since the luminance channel comprises only one baseband component in the frequency domain, the luminance gain in the amplitude of this component, i.e.,

$$\begin{aligned}\gamma_l &= \frac{1}{L}||\widetilde{\boldsymbol{h}}_l||_2 = \frac{1}{L}\sqrt{\widetilde{h}_l^2(0,0) + 0 + \ldots + 0} \\ &= \frac{1}{L}\widetilde{h}_l(0,0).\end{aligned}$$

Substituting in the DFT equation with $u = v = 0$, we get

$$\begin{aligned}\gamma_l(\boldsymbol{x}) &= \frac{1}{L}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} h_l(m,n) \\ &= \frac{1}{L}\mathbf{1}^T\boldsymbol{h}_l = \frac{1}{L}\mathbf{1}^T\boldsymbol{Z}_l\boldsymbol{x} = \boldsymbol{b}^T\boldsymbol{x},\end{aligned}$$

where $\boldsymbol{b} \stackrel{\text{def}}{=} \frac{1}{L}\mathbf{1}^T\boldsymbol{Z}_l$. As for the chrominance gain γ_c, by squaring the definition in (5.17), we get

$$\begin{aligned}\gamma_c(\boldsymbol{x})^2 &= \frac{1}{L^2}\min\left(||\widetilde{\boldsymbol{h}}_\alpha||_2^2, ||\widetilde{\boldsymbol{h}}_\beta||_2^2\right) \\ &\stackrel{(a)}{=} \min\left(||\boldsymbol{h}_\alpha||_2^2, ||\boldsymbol{h}_\beta||_2^2\right) \\ &= \min\left(||\boldsymbol{Z}_\alpha x||_2^2, ||\boldsymbol{Z}_\beta x||_2^2\right) = \min\left(\boldsymbol{x}^T\boldsymbol{Q}_\alpha\boldsymbol{x}, \boldsymbol{x}^T\boldsymbol{Q}_\beta\boldsymbol{x}\right),\end{aligned} \tag{C.16}$$

where (a) follows from Parseval theorem, and $\boldsymbol{Q}_\alpha \stackrel{\text{def}}{=} \boldsymbol{Z}_\alpha^T\boldsymbol{Z}_\alpha$ and $\boldsymbol{Q}_\beta \stackrel{\text{def}}{=} \boldsymbol{Z}_\beta^T\boldsymbol{Z}_\beta$ are two positive semidefinite matrices.

VITA

VITA

Omar A. Elgendy received the B.Sc. degree in Electronics and Communications Engineering (Distinction with honors) in 2010 and the M.Sc. degree in Engineering Mathematic in 2015, from Faculty of Engineering, Cairo University. He is currently a PhD student and a Research Fellow in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. He has been working as an RA in Statistical Signal and Image Processing (SSIP) lab since August 2015 under the supervision of Prof. Stanley H. Chan. His research interests include statistical signal processing, with applications to single-photon imaging, low-light imaging, HDR imaging, image reconstruction algorithms, imaging on mobile devices, and denoising. In addition, he has an experience in statistical pattern classification, with applications to speaker recognition; wireless communications; and optimization algorithms for large-scale optimization problems.

Elgendy was the recipient of the Best Paper Award of IEEE International Conference on Image Processing (ICIP) in 2016 for his paper (with Stanley H. Chan) on the image reconstruction and threshold design for Quanta Image Sensor (QIS). He is also the recipient of the Bilsland Dissertation Fellowship for the academic year 2018/2019 from Purdue university.