ADVANCING THE DIGITAL CAMERA PIPELINE FOR MOBILE MULTIMEDIA: KEY CHALLENGES FROM A SIGNAL PROCESSING PERSPECTIVE

Keigo Hirakawa and Patrick J. Wolfe

Harvard University, School of Engineering and Applied Sciences
Oxford Street, Cambridge, MA 02138 USA
{hirakawa@stat.harvard.edu, patrick@seas.harvard.edu}

ABSTRACT

The ubiquity of digital color image content continues to raise consumer technological awareness and expectations, and places a greater demand than ever on algorithms that support color image acquisition for mobile devices. In this paper we consider the key signal processing challenges to advancing the digital camera pipeline for mobile multimedia, with a particular focus on advances that have the potential to enhance image quality and reduce overall cost and power consumption. We first examine key technical challenges to pipeline design presented by demands such as shrinking device footprints, increasing throughput, and enhancing color fidelity. We then describe a recently introduced analytical framework based on spatio-spectral sampling for color image acquisition, and discuss its potential implications for quality and cost improvements. We then describe a number of resolution-distortion trade-offs, in particular noise processes and crosstalk, and show via simulation how a spatio-spectral acquisition framework helps to pinpoint aspects of pipeline design that can enhance computational efficiency and performance simultaneously.

Index Terms— image sensors, image sampling, image reconstruction, image denoising, image color analysis

1. INTRODUCTION

With the annual sales of mobile phones projected to exceed one billion handsets by 2009, mobile multimedia is well positioned to become a mainstream platform for entertainment. Despite narrowing profit margins and increased competition, the growing ubiquity of digital multimedia drives the demands for increased throughput and improved image quality in color imaging devices. Low-cost and low-power hardware designs nevertheless top the list of priorities for the mobile phone industry, and in this respect signal-processing solutions offer attractive benefits and cost savings that cannot be ignored.

In this paper we develop a signal processing perspective on the future of the digital camera pipeline, with a particular focus on the advances that have the potential to enhance image quality and reduce overall cost and power consumption. Our interest lies in quantifying the trade-offs between performance and complexity—and we do so not by explicit comparisons of digital camera pipelines, but by considering the imaging system's resilience to noise, aliasing, and artifacts that make the subsequent data processing steps more complicated and expensive. Given that mushrooming data rates pose additional computational, transmission, and storage challenges that cannot be solved by the advances in hardware alone, this type of analysis presents opportunities for new contributions in low-complexity low-cost digital camera designs through signal processing advancements.

For example, the inherent shortcomings of color filter array designs mean that subsequent processing steps often yield diminishing returns in terms of image quality, and in our previous work we proposed a novel acquisition scheme that preserves the integrity of the signal during acquisition [1]. Shrinking the device footprint is another important step toward increasing the pixel sensor count in a cost-effective manner—a trend partly fueled by the popular perception that higher spatial resolution necessarily leads to better image quality. The device footprint reduction problem is complicated by the problems such as noise and crosstalk that are difficult to model and quantify [2–5]. As the image sensor represents the first step in the digital camera pipeline, it largely determines the image quality achievable by subsequent processing schemes.

In this paper we offer a signal processing framework for understanding the resolution-distortion trade-offs, its implications for the complexity of the subsequent processing steps, and the possibilities for future improvements. Our analysis extends the spatio-spectral sampling theory for color image acquisition that provides insight into the trade-offs between the effective resolution and degree of degradation due to aliasing [1]. In particular, we present a signal processing perspective on the components of the digital camera pipeline, such as sampling, noise, crosstalk, and reconstruction, and evaluate its expected performance using the prior knowledge of image signals.

2. BACKGROUND

2.1. Review: Color Image Sensor

In this section, we review components of the image sensor that play important roles in determining overall limits in achievable image quality relative to subsequent processing steps. Let $\boldsymbol{x}(\boldsymbol{t}) = [x_r(\boldsymbol{t}), x_g(\boldsymbol{t}), x_b(\boldsymbol{t})]$ be the RGB tristimulus value of the desired *continuous* color image signal at spatial location $\boldsymbol{t} \in \mathbb{R}^2$. Each pixel sensor is equipped with a microlens, an optical MEMS device designed to increase the *fill factor* (area of the spatial integration) by locally focusing the light away from circuitries and toward the regions of the pixel sensor that are photosensitive. Given a spatial sampling interval τ for the sensor, the integrated light $\boldsymbol{x}(\boldsymbol{n}) = [x_r(\boldsymbol{n}), x_g(\boldsymbol{n}), x_b(\boldsymbol{n})]$ at pixel location $\boldsymbol{n} \in \mathbb{Z}^2$ is:

$$\boldsymbol{x}(\boldsymbol{n}) = \begin{bmatrix} \{h * x_r\}(\tau\boldsymbol{n}) \\ \{h * x_g\}(\tau\boldsymbol{n}) \\ \{h * x_b\}(\tau\boldsymbol{n}) \end{bmatrix}, \tag{1}$$

where '$*$' indicates convolution and $h(\boldsymbol{t})$ is a filter that represents a spatial integration over the pixel sensor. The photons collected by the microlens must then penetrate through color filter before reaching the photosensitive element of the sensor. CMOS photo diode active pixel sensors measure the intensity of the light using a photo diode and three transistors, all major sources of noise [6]. CCD sensors, on the other hand, rely on the electron-hole pair that is generated when a photon strikes silicon [7].

A color filter array (CFA) is a physical construction whereby the spectral components of the light are spatially multiplexed—that is, each pixel location measures the intensity of the light corresponding to only a single color [8]. Let $\boldsymbol{c}(\boldsymbol{n}) = [c_r(\boldsymbol{n}), c_g(\boldsymbol{n}), c_b(\boldsymbol{n})]$ represent the CFA color combination corresponding to $\boldsymbol{x}(\boldsymbol{n})$, and d_r and d_b be the DC components of c_r and c_b. If $c_r + c_g + c_b = \lambda$ for some constant λ, then light penetrating the color filter may be written as:

$$y(\boldsymbol{n}) = \boldsymbol{c}(\boldsymbol{n})^T \boldsymbol{x}(\boldsymbol{n}) = \lambda[c_\alpha(\boldsymbol{n}), 1, c_\beta(\boldsymbol{n})] \begin{bmatrix} x_\alpha(\boldsymbol{n}) \\ x_\ell(\boldsymbol{n}) \\ x_\beta(\boldsymbol{n}) \end{bmatrix}, \tag{2}$$

where $x_\alpha = x_r - x_g$, $x_\beta = x_b - x_g$ are difference images, $x_\ell = x_g + d_r x_\alpha + d_b x_\beta$ is a baseband component, and $c_\alpha = c_r - d_r$ and $c_\beta = c_b - d_b$ are modulation carrier frequencies [1]. The advantage of the $\{x_\alpha, x_\ell, x_\beta\}$ representation is the difference images enjoy rapid spectral decay away from their center frequencies, whereas baseband copy x_ℓ embodies the edge and texture information; moreover, $\{x_\alpha, x_\ell, x_\beta\}$ are generally observed to be only weakly correlated [9].

While a detailed investigation of noise sources is beyond the scope of this paper, studies suggest that $z(\boldsymbol{n})$, the number of photons encountered during a spatio-temporal integration, is a Poisson process denoted as $z(\boldsymbol{n})|y(\boldsymbol{n}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}(k \cdot y(\boldsymbol{n}))$, where k is a proportionality constant that scales linearly with the integration time and surface area of pixels and lens. Note $E[z|y] = ky$, $\text{Var}(z|y) = ky$, and when ky sufficiently large, $p(z|y)$ converges weakly to the normal distribution $\mathcal{N}\big(ky, ky\big)$. In practice, the photo diode charge (e.g. photodetector readout signal) is assumed proportional to $z(\boldsymbol{n})$, thus we interpret $ky(\boldsymbol{n})$ and $z(\boldsymbol{n})$ as the ideal and noisy sensor data at pixel location $\boldsymbol{n}$, respectively.

2.2. Review: Digital Camera Pipeline

Given the sensor data $z(\boldsymbol{n})$, the goal of the digital camera pipeline is to estimate the color image $\boldsymbol{x}(\boldsymbol{t})$—note that we use the *continuous* representation of the image here in order to better compare images captured at different sampling rates. As stated earlier, one cost-effective measure to increase spatial resolution is simply to shrink the device footprint in hardware. However, device physics and simple geometric arguments (fewer incident photons, for example) dictate that this increase in spatial resolution will be accompanied by a corresponding increase in sensor noise effects. To understand this trade-off, we first review a number of signal processing steps or modules that comprise a camera pipeline after the acquisition of data:

- The spatial subsampling due to the implementation of color filter array is approximately inverted through **demosaicking**—yielding a complete tristimulus value at each pixel location. Assuming y (ideal sensor data) as an input to the demosaicking algorithm, demosaicking is a demultiplexing of frequency multiplexed signals $\{x_\alpha, x_\ell, x_\beta\}$ [1, 10].
- Because the color coordinates defined by the sensitivity of the color filters $\boldsymbol{c}$ may not correspond exactly to the standardized color space (such as sRGB space), the resulting tristimulus values undergo a **color space conversion** (change of basis) via pixel-wise multiplication by a predetermined matrix $\boldsymbol{M} \in \mathbb{R}^{3\times 3}$, $\boldsymbol{x}_0(\boldsymbol{n}) = \boldsymbol{M}_0 \boldsymbol{x}(\boldsymbol{n})$. Additional color space conversion may be required for image compression, which usually operates in an opponent color space.
- Given the variability of the Poisson process, the **Poisson mean** y is often inaccessible and must be estimated from z. Alternatively, demosaicking methods applied to z as a proxy for y yield a "noisy" estimate of $\boldsymbol{x}$.
- Human visual systems make adjustments to the color to account for variations in illuminant and the environment. Linear **white-balance correction**—needed to match the camera output with the *perceived* color—is a function of the estimated scene illuminant and typically takes place either before demosaicking or concurrently with the color space conversion: $\boldsymbol{x}_1(\boldsymbol{n}) = \boldsymbol{M}_1(\text{illuminant})\boldsymbol{x}(\boldsymbol{n})$.
- A point-wise nonlinearity termed the **inverse gamma function** $\Gamma^{-1} : \mathbb{R} \to \mathbb{R}$ is applied to the color-corrected tristimulus value $\boldsymbol{x}$ to yield the display stimulus $\boldsymbol{u} = [u_1, u_2, u_3]^T$, $u_i(\boldsymbol{n}) = \Gamma^{-1}\{x_i(\boldsymbol{n})\}$. This gamma correction step will undo the effects of nonlinearity Γ inherent in display devices (i.e., $\Gamma(\boldsymbol{u})$ is linear with respect to $\boldsymbol{x}$).

Each of the key components in a typical camera pipeline is aimed at correcting or enhancing certain aspects of the hardware or human-hardware interface—and in particular, the demosaicking, color space conversion, and Poisson mean estimation steps are explicitly coupled to the image data acquisition process highlighted in the previous section. The key challenges for advancing the digital camera pipeline from a signal processing perspective, therefore, involve resolution-distortion trade-offs as manifested by the color image sensor itself—the subject of the rest of this article.

3. RESOLUTION-DISTORTION TRADE-OFFS

In this section, we examine the inherent trade-offs between spatial resolution and signal-dependent measurement noise and other distortions via spatio-spectral sampling theory [1]. The main result of our analysis is that noise dependency is increasingly severe as sensor size decreases.

3.1. Resolution and Poisson Process

For analytical tractability, let $h(\boldsymbol{t})$ be an ideal low-pass filter. Combining (1), (2), and the effects of the Poisson process, the measurement $z(\boldsymbol{n})$ can be characterized as:

$$\begin{aligned} E[z(\boldsymbol{n})|\boldsymbol{x}] &= \text{Var}(z(\boldsymbol{n})|\boldsymbol{x}) \qquad (3) \\ &= k\lambda\{h * x_\ell\}(\tau\boldsymbol{n}) + k\nu_h\lambda\Big(c_\alpha(\boldsymbol{n})x_\alpha(\tau\boldsymbol{n}) + c_\beta(\boldsymbol{n})x_\beta(\tau\boldsymbol{n})\Big), \end{aligned}$$

where ν_h is the DC component of the convolution filter $h(\boldsymbol{t})$ and for sufficiently low-bandwidth difference images, $h * x_\alpha = \nu_h x_\alpha$ and $h * x_\beta = \nu_h x_\beta$. From the spatio-spectral sampling perspective, $z(\boldsymbol{n})$ is a lowpass version of x_ℓ (i.e., $k\lambda\{h * x_\ell\}(\tau\boldsymbol{n})$) that has been corrupted by two sources of degradation: aliasing (i.e., $k\lambda c_\alpha(\boldsymbol{n})\{h * x_\alpha\}(\tau\boldsymbol{n}) + k\lambda c_\beta(\boldsymbol{n})\{h * x_\beta\}(\tau\boldsymbol{n})$), and the variability resulting from the Poisson process (i.e. $\text{Var}(z(\boldsymbol{n})|\boldsymbol{x})$). Reconstruction amounts to separating x_ℓ from these interfering signals in $z(\boldsymbol{n})$—without making any additional assumptions, such as the local sparsity assumed by contemporary nonlinear demosaicking algorithms [9]. A little algebra will verify that the *distortion* in $z(\boldsymbol{n})$ relative to $x_\ell(\boldsymbol{t})$ can be measured as:

$$J(\boldsymbol{x}) = \left\| \frac{\mathcal{W}\{z\}(\boldsymbol{t})}{k\nu_h\lambda} - x_\ell(\boldsymbol{t}) \right\|^2,$$

where $\mathcal{W}\{z\}$ is the Whittaker-Shannon ideal reconstruction of the discrete samples $z(\boldsymbol{n})$ (i.e. orthogonal projection to space of bandlimited functions). The expectation $E[J(\boldsymbol{x})]$ may further be decomposed as

$$E\left[\left\|\frac{\mathcal{W}\{y\}(\boldsymbol{t})}{\nu_h\lambda} - x_\ell(\boldsymbol{t})\right\|^2 + \left\|\frac{\mathcal{W}\{z-ky\}(\boldsymbol{t})}{k\nu_h\lambda}\right\|^2\right]. \qquad (4)$$

Let $E[x_\alpha] = E[x_\beta] = 0$ and $\{x_\alpha, x_\ell, x_\beta\}$ be mutually independent. Then the first term in (4) expands to:

$$E[\|\{h * x_\ell\}/\nu_h - x_\ell\|^2 + \|c_\alpha x_\alpha + c_\beta x_\beta\|^2]. \qquad (5)$$

The first term in (5) is the degradation due to loss of resolution, and it is independent of the choice of color filter array and the surface area of the pixel sensor. The second term in (5) is the aliasing from CFA sampling, which is independent of the surface area of the pixel sensor and the resolution. Similarly, the second term in (4) is the measurement noise; from (3), we see that this is equivalent to

$$E[\mathrm{Var}(z(\boldsymbol{n})|\boldsymbol{x})/(k\nu_h\lambda)^2] = \{h * x_\ell\}(\tau\boldsymbol{n})/(k\nu_h^2\lambda). \qquad (6)$$

The conclusion we draw from the above exercise is that larger pixel sensor area (i.e. larger k) and panchromatic CFA pattern (i.e. larger λ) are favorable for reducing the measurement noise. The effects of the resolution on noise (h and ν_h) are signal dependent, though k and h are often inversely coupled, and the trade-off between (5) and (6) is not straightforward. The expected distortion $E[J(\boldsymbol{x})]$ can be evaluated empirically using simulation—choosing widely available test images for $\boldsymbol{x}$, Figure 1 shows distortion as a function of sensor resolution. A fixed value of k means that the overall image sensor size and integration time is held constant while the sensor surface is divided up into smaller pixels to increase resolution. The key observation here is that despite increased resolution, shrinking pixel sensors may result in more distortion in some cases. The problem is especially bad when k or $\boldsymbol{x}$ is small (small sensor size, small lens, low-light environment, etc). The graph also suggests that a better CFA design reduces distortion far more effectively than increasing the pixel count.

3.2. Resolution and Crosstalk

Crosstalk, a phenomenon where photon or electron leakages cause an interaction between neighboring pixels, is a major problem when the device footprint decreases because of reduced distances between pixel sensors. Two major contributions to crosstalk we consider here are optical diffraction and minority carrier diffusion.

Optical diffraction occurs when a high incidence angle of the light entering the substrate causes the photons to stray away from the center of the pixel; microlenses can help to reduce this risk [2]. The diffusion is stochastic but mostly linear with respect to the intensity of the light. The incident angle is typically wider for the pixel sensors far from the lens axis, and thus the light that reaches photosensitive material can be modeled as *spatially-variant* convolution: $\hat{y}(\boldsymbol{n}) = \sum_{\boldsymbol{m}} y(\boldsymbol{m}) f(\boldsymbol{n}, \boldsymbol{m})$ where $f(\boldsymbol{n}, \boldsymbol{m})$ is the location-dependent impulse response. The precise modeling of $f(\boldsymbol{n}, \boldsymbol{m})$ as a function of sensor geometry is an active area of research involving sophisticated simulation [2, 4]. Nevertheless, location-*independent* approximation of the point-spread-function f using ideal low-pass filters have been suggested [5]. As the coupling between pixels occur before charge collection, the Poisson noise is spatially uncorrelated, so that $z(\boldsymbol{n}) \overset{\text{i.i.d}}{\sim} \mathcal{P}(\hat{y}(\boldsymbol{n}))$.

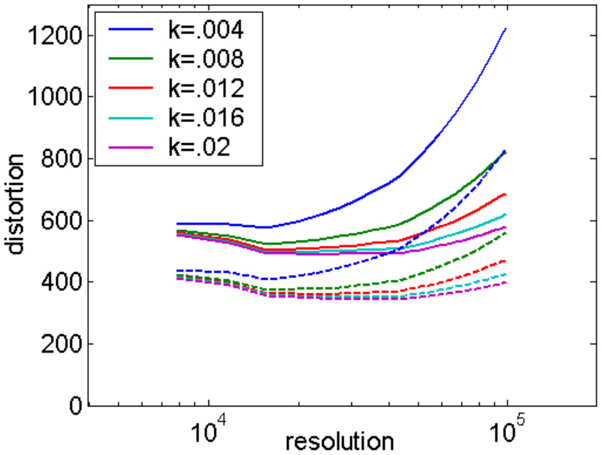

Fig. 1. Distortion with respect to x_ℓ as a function of the sensor resolution while holding the sensor size constant, measured from a simulation using 24-bit 512×768 images. Solid lines indicate Bayer CFA [8], dashed lines indicate spatio-spectral CFA design [1].

Minority carrier diffusion deteriorates the signal when photons stray from the target after the charge is collected [3]. This carrier is typically deterministic and mostly linear with respect to the signal strength, and it can be modeled as *spatially-invariant* convolution: $\hat{z}(\boldsymbol{n}) = \sum_{\boldsymbol{m}} z(\boldsymbol{n} - \boldsymbol{m}) g(\boldsymbol{m})$, where $g(\boldsymbol{m})$ is the convolution kernel. Note that the Poisson noise in $\hat{z}$ is no longer spatially uncorrelated. Motivated by physics, the characteristics of this diffusion process are crudely modeled as $g(\boldsymbol{n}) \propto e^{-\|\tau\boldsymbol{n}\|/L}$, where L is the diffusion constant and τ is the sample interval [3].

Using the updated definitions $\hat{y}$ and $\hat{z}$, distortion with crosstalk is measured as:

$$\hat{J}(\boldsymbol{x}) = \left\|\frac{\mathcal{W}\{\hat{z}\}(\boldsymbol{t})}{k\nu_f\nu_g\nu_h\lambda} - x_\ell(\boldsymbol{t})\right\|^2,$$

where ν_f and ν_g are the DC values of the convolution filters f and g. Breaking down into loss of resolution, aliasing, and noise as before, $E[\hat{J}(\boldsymbol{x})]$ is equivalent to the sum of the following terms:

$$\begin{aligned} &E[\|\{f * g * h * x_\ell\}/(\nu_f\nu_g\nu_h) - x_\ell\|^2], \\ &E[\|f * g * \{c_\alpha x_\alpha + c_\beta x_\beta\}\|^2], \\ &E[g^2 * \{f * h * x_\ell\}(\tau\boldsymbol{n})/(k\nu_f^2\nu_g^2\nu_h^2\lambda)]. \end{aligned} \qquad (7)$$

As before, Figure 2 evaluates the expected distortion $E[\hat{J}(\boldsymbol{x})]$ empirically using test images. A rather surprising consequence of (7) is that the low-pass convolution filters g and f may help suppress the distortion in $\hat{z}(\boldsymbol{n})$ relative to $x_\ell(\boldsymbol{t})$ because they attenuate the aliasing components ($c_\alpha x_\alpha + c_\beta x_\beta$), which, owing to the carrier frequencies c_α and c_β, occupy the high-pass region.

The real penalty imposed by crosstalk in the trade-off analysis is the reconstructibility of difference images x_α and x_β, which roughly correspond to the chrominance of the image. The reconstruction of x_α and x_β depends greatly on the preservation of modulated signal $c_\alpha x_\alpha + c_\beta x_\beta$. Consider, for example, the amplitude response of g at the highest modulation frequencies in c_α and c_β as a function of resolution (assuming a fixed overall sensor area). Owing to the rapid spectral decay of Gaussian filters, the modulated signal

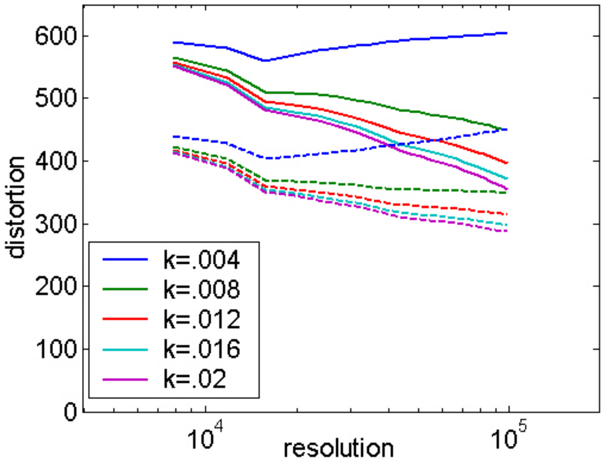

Fig. 2. Distortion with respect to x_ℓ as a function of the sensor resolution with crosstalk artifacts. Solid lines indicate Bayer CFA [8], dashed lines indicate spatio-spectral CFA design [1].

$c_\alpha x_\alpha + c_\beta x_\beta$ is attenuated very quickly as the pixel sensor geometry shrinks. Another important observation is that crosstalk problems persist regardless of illuminant or noise level, as convolution filters f and g are linear.

The conclusion we draw from the above is that due to attenuation of chrominance information, crosstalk results in desaturation of color and increased sensitivity to noise. This confirms our intuition that photon and electron leakage from neighboring pixels results in linearly combining measurements from different color filters, thereby deteriorating the quality of information pertaining to color. Moreover, the analysis in (7) informs us that the estimation of x_α and x_β—formulated as inverse crosstalk problem—would involve properly scaling the chrominance by the inverse of the amplitude response of f and g at the modulation frequencies induced by a particular CFA pattern.

4. DISCUSSION AND CONCLUSION

Motivated by the perspective that noise, aliasing, and artifacts in an imaging system lead to more complicated and expensive signal processing steps in digital camera pipeline, we have offered here a signal processing perspective on trade-offs between resolution and distortion as device footprints continue to shrink. We characterized the color image sensor in terms of physical properties such as spatio-temporal integration, color filter array, Poisson process, and electron/photon leakage, and analytically and numerically evaluated the distortion in the measured sensor data. We found that advantages to shrinking pixel sensor geometries as a means to increase resolution in a cost-effective manner may be overridden by Poisson noise in the signal measurement process, and that better CFA designs have the potential to reduce distortion far more effectively. Our analysis of resolution-crosstalk trade-offs revealed the mechanism by which crosstalk desaturates the colors while sometimes improving estimates for the luminance component.

5. REFERENCES

[1] K. Hirakawa and P. J. Wolfe, "Spatio-spectral color filter array for enhanced image fidelity," in *IEEE International Conference on Image Processing*, 2007, vol. 2, pp. 81–84.

[2] G. Agranov, V. Berezin, and R. H. Tsai, "Crosstalk and microlens study in a color CMOS image sensor," *IEEE Transactions on Electron Devices*, vol. 50, no. 1, pp. 4–11, 2003.

[3] I. Shcherback, T. Danov, and O. Yadid-Pecht, "A comprehensive CMOS APS crosstalk study: Photoresponse model, technology, and design trends," *IEEE Trans. Electron Devices*, vol. 51, no. 21, pp. 2033–2041, 2004.

[4] H. Rhodes, G. Agranov, C. Hong, U. Boettiger, R. Mauritzson, J. Ladd, I. Karasev, J. McKee, E. Jenkins, W. Quinlin, I. Patrick, J. Li, X. Fan, R. Panicacci, S. Smith, C. Mouli, and J. Bruce, "CMOS imager technology shrinks and image performance," in *IEEE Workshop on Microelectronics and Electron Devices*, 2004, pp. 7–18.

[5] T. Q. Pham, L. J. van Vliet, and K Schutte, "Influence of signal-to-noise ratio and point spread function on limits of super-resolution," in *SPIE-IS&T Electronic Imaging: Algorithms and Systems IV*, 2005, pp. 169–180.

[6] H. Tian, B. Fowler, and A. E. Gamal, "Analysis of temporal noise in CMOS photodiode active pixel sensor," *IEEE Journal of Solid State Circuits*, vol. 36, no. 1, pp. 92–101, January 2001.

[7] G. E. Healey and R. Kondepudy, "Radiometric CCD camera calibration and noise estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 3, pp. 267–276, March 1994.

[8] B. E. Bayer, "Color imaging array," US Patent 3 971 065, 1976.

[9] B. K. Gunturk, J. Glotzbach, Y. Altunbasak, R. W. Schafer, and R. M. Mersereau, "Demosaicking: Color filter array interpolation in single chip digital cameras," *IEEE Signal Processing Magazine*, vol. 22, no. 1, pp. 44–54, January 2005.

[10] E. Dubois, "Filter design for adaptive frequency-domain Bayer demosaicking," *Proceedings of the IEEE International Conference on Image Processing*, pp. 2705–2708, 2006.