FlatCam: Thin, Bare-Sensor Cameras using Coded Aperture and Computation

M. Salman Asif,[1] Ali Ayremlou,[1] Aswin Sankaranarayanan,[2] Ashok Veeraraghavan,[1] and Richard Baraniuk[1]
[1]Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, USA
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

FlatCam is a thin form-factor lensless camera that consists of a coded mask placed on top of a bare, conventional sensor array. Unlike a traditional, lens-based camera where an image of the scene is directly recorded on the sensor pixels, each pixel in FlatCam records a linear combination of light from multiple scene elements. A computational algorithm is then used to demultiplex the recorded measurements and reconstruct an image of the scene. FlatCam is an instance of a coded aperture imaging system; however, unlike the vast majority of related work, we place the coded mask extremely close to the image sensor that can enable a thin system. We employ a separable mask to ensure that both calibration and image reconstruction are scalable in terms of memory requirements and computational complexity. We demonstrate the potential of the FlatCam design using two prototypes: one at visible wavelengths and one at infrared wavelengths.

I. Introduction

A range of new imaging applications is driving the miniaturization of cameras. As a consequence, significant progress has been made towards minimizing the total volume of the camera, which has enabled new applications in endoscopy, pill cameras, and in vivo microscopy. Unfortunately, this strategy of miniaturization has an important shortcoming: the amount of light collected at the sensor decreases dramatically as the lens aperture and the sensor size become smaller. As a consequence, ultra-miniature imagers built simply by scaling down the optics and sensors suffer from extremely low light collection.

In this paper, we present a camera architecture that we call FlatCam, which is inspired by coded aperture imaging principles pioneered in astronomical x-ray and gamma-ray imaging [1]–[4]. Our proposed FlatCam design uses a very large photosensitive area with a very thin form factor. The FlatCam achieves thin form factor by dispensing with a lens and replacing it with a coded, binary mask placed almost immediately atop a bare conventional sensor array. The image formed on the sensor can be viewed as a superposition of many pinhole images. Thus, the light collection ability of such a coded aperture system is proportional to the size of the sensor and the transparent regions (pinholes) in the mask. In contrast, the light collection ability of a miniature, lens-based camera is limited by the lens aperture size, which is restricted by the requirements on the device thickness.

An illustration of the FlatCam design is presented in Fig. 1. Light from a scene passes through a coded mask and lands on a conventional image sensor. The mask consists of opaque and transparent features (to block and transmit light, respectively); each transparent feature can be viewed as a pinhole. Light from the scene gets diffracted by the mask features such that light from each scene location casts a unique mask shadow on the sensor, and this mapping can be represented using a linear operator. A computational algorithm then inverts this linear operator to recover the original light distribution of the scene from the sensor measurements.

Corresponding author: M. Salman Asif (email: sasif@rice.edu).

Our FlatCam design has many attractive properties besides its slim profile. First, since it reduces the thickness of the camera but not the area of the sensor, it collects more light than miniature, lens-based cameras with same thickness. Second, the mask can be created from inexpensive materials that operate over a broad range of wavelengths. Third, the mask can be fabricated simultaneously with the sensor array, creating new manufacturing efficiencies. The mask can be fabricated either directly in one of the metal interconnect layers on top of the photosensitive layer or on a separate wafer thermal compression that is bonded to the back side of the sensor, as is typical for back-side illuminated image sensors [5].

We demonstrate the potential of the FlatCam using two prototypes built in our laboratory with commercially available sensors and masks: a visible prototype in which the mask-sensor spacing is about 0.5mm and a short-wave infrared (SWIR) prototype in which the spacing is about 5mm. Figures 4 and 8 illustrate sensor measurements and reconstructed images using our prototype FlatCams.

II. Related work

Pinhole cameras. Imaging without a lens is not a new idea. Pinhole cameras, the progenitor of lens-based cameras, have been well known since Alhazen (965–1039AD) and Mozi (c. 370BCE). However, a tiny pinhole drastically reduces the amount of light reaching the sensor, resulting in noisy, low-quality images. Indeed, lenses were introduced into cameras for precisely the purpose of increasing the size of the aperture, and thus the light throughput, without degrading the sharpness of the acquired image.

Coded apertures. Coded aperture cameras extend the idea of a pinhole camera by using masks with multiple pinholes [1]–[3]. The primary goal of coded aperture cameras is to increase the light throughput compared to a pinhole camera. Figure 2 summarizes some salient features of pinhole, lens-based, and FlatCam (coded mask-based) architectures.

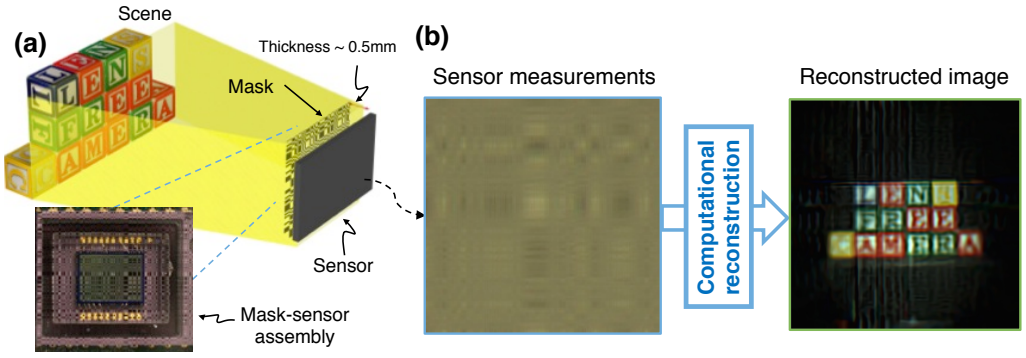


Fig. 1: FlatCam architecture. **(a)** Every light source within the camera field-of-view contributes to every pixel in the multiplexed image formed on the sensor. A computational algorithm reconstructs the image of the scene. Inset shows the mask-sensor assembly of our prototype in which a binary, coded mask is placed 0.5mm away from an off-the-shelf digital image sensor. **(b)** An example of sensor measurements and the image reconstructed by solving a computational inverse problem.

Coded-aperture cameras have traditionally been used for imaging wavelengths beyond the visible spectrum (e.g., x-ray and gamma-ray imaging), for which lenses or mirrors are expensive or infeasible [1]–[4], [6]. In recent years, coded aperture-based systems using compressive sensing principles [7]–[9] have been studied for image super-resolution [10], spectral imaging [11], and video capture [12]. Mask-based lens-free designs have also been proposed for flexible field-of-view selection in [13], compressive single-pixel imaging using a transmissive LCD panel [14], and for separable coded masks [15].

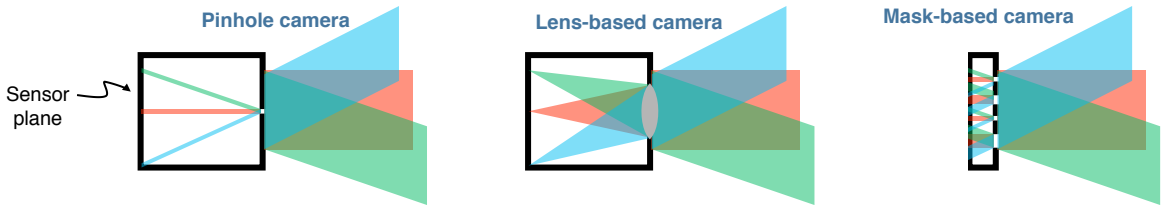
Existing coded aperture-based lensless systems have two main limitations: First, the large body of work devoted to coded apertures invariably place the mask significantly far away from the sensor (e.g., 65mm distance in [15]). In contrast, our FlatCam design offers a thin form factor. For instance, in our prototype with a visible sensor, the spacing between the sensor and the mask is only 0.5mm. Second, the masks employed in some designs have transparent features only in a small central region whose area is invariably much smaller than the area of the sensor. In contrast, almost half of the features (spread across the entire surface) in our mask are transparent. As a consequence, the light throughput of our designs are many orders of magnitude larger as compared to previous designs. Furthermore, the lensless cameras proposed in [14], [15] use programmable spatial light modulators (SLM) and capture multiple images while changing the mask patterns. In contrast, we use a static mask in our design, which can potentially be fixed on the sensor during fabrication or the assembly process.

Camera arrays. A number of thin imaging systems have been developed over the last few decades. The TOMBO architecture [16], inspired by insect compound eyes, reduces the camera thickness by replacing a single, large focal-length lens with multiple, small focal-length microlenses. Each microlens and the sensor area underneath it can be viewed as a separate low-resolution, lens-based camera, and a single high-resolution image can be computationally reconstructed by fusing all of the sensor measurements. Similar architectures have been used for designing thin infrared cameras [17]. The camera thickness in this design is dictated by the geometry of the microlenses; reducing the camera thickness requires a proportional reduction in the sizes of the microlenses and sensor pixels. As a result, microlens-based cameras currently offer only up to a four-fold reduction in the camera thickness [18], [19].

Folded optics. An alternate approach for achieving thin form factors relies on folded optics, where light manipulation similar to that of a traditional lens is achieved using multi-fold reflective optics [20]. However, folded optics based systems have low light collection efficiencies.

Ultra-miniature lensless imaging with diffraction gratings. Recently, miniature cameras with integrated diffraction gratings and CMOS image sensors have been developed [21]–[24]. These cameras have been successfully demonstrated on tasks such as motion estimation and face detection. While these cameras are indeed ultra-miniature in total volume (100 micron sensor width by 200 micron thickness), they retain the large thickness-to-width ratio of conventional lens-based cameras. Because of the small sensor size, they suffer from reduced light collection ability. In contrast, in our visible prototype below, we used a 6.7mm wide square sensor, which increases the amount of light collection by about three orders of magnitude, while the device thickness remains approximately similar (500 micron).

Lensfree microscopy and shadow imaging. Lensfree cameras have been successfully demonstrated for several microscopy and lab-on chip application, wherein the subject to be imaged is close to the image sensor. An on-chip, lens-free microscopy design that uses amplitude masks to cast a shadow of point illumination sources onto a microscopic tissue sample has shown significant promise for microscopy and related applications, where the sample being imaged is very close to the sensor (less than 1mm) [25], [26]. Unfortunately, this technique cannot be directly extended to traditional photography and other applications that require larger standoff distances and do not provide control over illumination.



	Pinhole camera	Lens-based camera	Mask-based camera
Light throughput	Low (~F/22)	High (~F/1.8)	High (~F/2.54)
Fabrication	Requires post-fabrication	Requires post-fabrication	Direct fabrication
Thickness	~10-20mm	~10-20mm	~10-500μm
Infrared and thermal	No wavelength limitations	Costly optics	No wavelength limitations
Cost	$	$$	$
Curved or flexible geometries	Infeasible due to limited field of view	Infeasible due to rigid optics	Adaptable to curved and flexible sensors

Fig. 2: Comparison of pinhole, lens-based, and coded mask-based cameras. Pinhole cameras and lens-based cameras provide one-to-one mapping between light from a focal plane and the sensor plane (note that light from three different directions is mapped to three distinct locations on the sensor), but the coded mask-based cameras provide a multiplexed image that must be resolved using computation. The table highlights some salient properties of the three camera designs. Pinholes cameras suffer from very low light throughput, while lens-based cameras are bulky and rigid because of their optics. In contrast, the FlatCam design offers thin, light-efficient cameras with the potential for direct fabrication.

III. FlatCam design

A. Replacing lenses with computation

Our FlatCam design places an amplitude mask almost immediately in front of the sensor array (see Fig. 1). While we focus on a single mask for exposition purposes, the concept extends to multiple amplitude masks in a straightforward manner. We assume that the sensor and the mask are planar, parallel to each other, and separated by distance d. For simplicity of explanation, we also assume (without loss of generality) that the mask modulates the impinging light in a binary fashion; that is, it consists of transparent features that transmit light and opaque features that block light. We denote the size of the transparent/opaque features by Δ and assume that the mask covers the entire sensor array.

Light from the scene is modulated and diffracted by the mask pattern and recorded on the image sensor. By assuming that the image formed on the sensor is a superposition of light sources in the scene, we can describe the transfer function between the scene image and the sensor measurements as

$$y = \Phi x + e. \tag{1}$$

Here, vector x denotes pixelated scene image, vector y denotes the sensor measurements, Φ denotes the transfer matrix, and e denotes the sensor noise and any model mismatch. Since the sensor pixels do not have a one-to-one mapping with the scene pixels, the matrix Φ will not resemble the identity matrix. Instead, each sensor pixel measures multiplexed light from multiple scene pixels, and each row of Φ indicates how strongly each scene pixel contributes to the intensity measured at a particular sensor pixel. In other words, any column in Φ denotes the image formed on the sensor if the scene contains a single, point light source at the respective location.

Multiplexing generally results in an ill-conditioned system. Our goal is to design a mask that produces a matrix Φ that is well conditioned and hence can be stably inverted without excessive noise amplification. We now discuss how we navigate among three inter-related design decisions: the placement d and feature size Δ of the mask, the mask pattern, and the image recovery (demultiplexing) algorithm.

B. Mask pattern

Three issues need to be considered in the design of the mask pattern: the light throughput, the complexity of system calibration and inversion, and the conditioning of the resulting multiplexing matrix Φ.

Light throughput. In the absence of the mask, the amount of light that can be sensed by the bare sensor is limited only by its CRA. Since the photosensitive element in a CMOS/CCD sensor array is situated in a small cavity, a micro-lens array directly on top of the sensor is used to increase the light collection efficiency. In spite of this, only light rays up to a certain angle of incidence reach the sensor, and this is the fundamental light collection limit of that sensor. Placing an amplitude-modulating mask very close to (and completely covering) the sensor results in a light-collection efficiency that is a fraction of the fundamental light collection limit of the sensor. In our designs, half of the binary mask features are transparent, which halves our light collection ability compared to the maximum limit. As described above, while it is true that the light collection ability of our FlatCam design is one-half of the maximum achievable with a particular sensor, the main advantage of the FlatCam design is that it allows us to use much larger sensor arrays for a given device thickness constraint, thereby significantly increasing the light collection capabilities of devices under thickness constraints.

Computational complexity. The (linear) relationship between the scene irradiance x and the sensor measurements y is contained in the multiplexing matrix Φ. Discretizing the

unknown scene irradiance into $N \times N$ pixel units and assuming an $M \times M$ sensor array, Φ is an $M^2 \times N^2$ matrix. Given a mask and sensor, we can obtain the entries of Φ either by modeling the transmission of light from the scene to the sensor or through a calibration process. Clearly, even for moderately sized systems, Φ is prohibitively large to either estimate (calibration) or invert (image reconstruction), in general. For example, to describe a system with a megapixel resolution scene and a megapixel sensor array, Φ will contain on the order of $10^6 \times 10^6 = 10^{12}$ elements.

One way to reduce the complexity of Φ is to use a separable mask for the FlatCam system. If the mask pattern is separable (i.e., an outer product of two one-dimensional patterns), then the imaging system in (1) can be rewritten as

$$Y = \Phi_L X \Phi_R^T + E, \tag{2}$$

where Φ_L, Φ_R denote matrices that correspond to one-dimensional convolution along the rows and columns of the scene, respectively, X is an $N \times N$ matrix containing the scene radiance, Y in an $M \times M$ matrix containing the sensor measurements, and E denotes the sensor noise and any model mismatch. For a megapixel scene and a megapixel sensor, Φ_L and Φ_R have only 10^6 elements each, as opposed to 10^{12} elements in Φ. Similar idea has been recently proposed in [15] with the design of doubly toeplitz mask. In our implementation, we also estimate the system matrices using a separate calibration procedure (see Sec. III-D), which also becomes significantly simpler for a separable system.

Conditioning. The mask pattern should be chosen to make the multiplexing matrices Φ_L and Φ_R as numerically stable as possible, which ensures a stable recovery of the image X from the sensor measurements Y. Such Φ_L and Φ_R should have low condition numbers, i.e., a flat singular value spectrum. For Toeplitz matrices, it is well known that, of all binary sequences, the so-called maximum length sequences, or M-sequences, have maximally flat spectral properties [27]. Therefore, we use a separable mask pattern that is the outer product of two one-dimensional M-sequence patterns. However, because of the inevitable non-idealities in our implementation, such as the limited sensor CRA and the larger than optimal sensor-mask distance due to the hot mirror, the actual Φ_L and Φ_R we obtain using a separable M-sequence based mask do not achieve a perfectly flat spectral profile. Nevertheless, as we demonstrate in our prototypes, the resulting multiplexing matrices enable stable image reconstruction in the presence of sensor noise and other non-idealities. All of the visible wavelength, color image results shown in this paper were obtained using normal, indoor ambient lighting and exposure times in 10–20ms range, demonstrating that robust reconstruction is possible.

C. Mask placement and feature size

The multiplexing matrices Φ_L, Φ_R describe the mapping of light emanating from the points in the scene to the pixels on the sensor. Consider light from a point source passing through one of the mask openings; its intensity distribution recorded at the sensor forms a point-spread function (PSF) that is due to both diffraction and geometric blurs. The PSF acts as a low-pass filter that limits the frequency content that can be recovered from the sensor measurements. The choice of the feature size and mask placement is dictated by the tradeoff between two factors: reducing the size of the PSF to minimize the total blur and enabling sufficient multiplexing to obtain a well-conditioned linear system.

The total size of the PSF depends on the diffraction and geometric blurs, which in turn depend on the distance between the sensor and the mask, d, and the mask feature size, Δ. The size of the diffraction blur is directly proportional to d and inversely proportional to Δ. The size of the geometric blur, however, is equal to the feature size Δ. This implies that the minimum blur radius is achieved when the two blur sizes are approximately equal. One possible way to reduce the size of the combined PSF is to move the mask closer to the sensor. However, the extent of multiplexing within the scene pixels shrinks as the mask moves closer to the sensor. Therefore, if we aim to keep the amount of multiplexing constant, then the mask feature size Δ should shrink proportionally to the mask-sensor distance d.

In practice, physical limits on the sensor-mask distance d or the mask feature size Δ can dictate the design choices. In our visible FlatCam prototype, for example, we use a Sony ICX285 sensor. The sensor has a 0.5mm thick hot mirror attached to the top of the sensor, which restricts the potential spacing between the mask and sensor surface. Therefore, we place the mask immediately atop the hot mirror, resulting in $d \approx 500\mu$m (distance between the mask and the sensor surface). We achieved the smallest total blur size using a mask feature size of approximately $\Delta = 30\mu$m. Of course, in future implementations, where the mask pattern is directly etched on top of the image sensor (direct fabrication) such a thickness constraint does not exist and we can achieve much higher resolution images by moving the mask closer to the sensor and reducing the mask feature size proportionally.

D. Camera calibration

We now provide the details of our calibration procedure for the separable imaging system modeled in (2). Instead of modeling the convolution shifts and diffraction effects for a particular mask-sensor arrangement, we directly estimate the system matrices.

To align the mask and sensor, we adjust their relative orientation such that a separable scene in front of the camera yields a separable image on the sensor. For a coarse alignment, we use a point light source, which projects a shadow of the mask onto the sensor, and align the horizontal and vertical edges on the sensor image with the image axes. For a fine alignment, we align the sensor with the mask while projecting horizontal and vertical stripes on a monitor or screen in the front of the camera.

To calibrate a system that can recover $N \times N$ images X, we estimate the left and right matrices Φ_L, Φ_R using the sensor measurements of $2N$ known calibration patterns projected on a screen as depicted in Fig. 3. Our calibration procedure relies

on an important observation. If the scene X is separable, i.e., $X = \mathbf{a}\mathbf{b}^T$ where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$, then

$$Y = \Phi_L \mathbf{a}\mathbf{b}^T \Phi_R^T = (\Phi_L \mathbf{a})(\Phi_R \mathbf{b})^T.$$

In essence, the image formed on the sensor is a rank-1 matrix, and by using a truncated singular value decomposition (SVD), we can obtain $\Phi_L \mathbf{a}$ and $\Phi_R \mathbf{b}$ up to a signed, scalar constant. We take N separable pattern measurements for calibrating each of Φ_L and Φ_R.

Specifically, to calibrate Φ_L, we capture N images $\{Y_1, \ldots, Y_N\}$ corresponding to the separable patterns $\{X_1, \ldots, X_N\}$ displayed on a monitor or screen. Each X_k is of the form $X_k = \mathbf{h}_k \mathbf{1}^T$, where $\mathbf{h}_k \in \mathbb{R}^N$ is a column of the orthogonal Hadamard matrix H of size $N \times N$ and $\mathbf{1}$ is an all-ones vector of length N. Since the Hadamard matrix consists of ± 1 entries, we record two images for each Hadamard pattern; one with $\mathbf{h}_k \mathbf{1}^T$ and one with $-\mathbf{h}_k \mathbf{1}^T$ while setting the negative entries to zero in both cases. We then subtract the two sensor images to obtain the measurements corresponding to X_k. Let $\widetilde{Y}_k = \mathbf{u}_k \mathbf{v}^T$ be the rank-1 approximation of the measurements Y_k obtained via SVD, where the underlying assumption is that $\mathbf{v} \approx \Phi_R \mathbf{1}$, upto a signed, scalar constant. Then, we have

$$[\mathbf{u}_1 \, \mathbf{u}_2 \, \cdots \mathbf{u}_N] = \Phi_L [\mathbf{h}_1 \, \mathbf{h}_2 \, \cdots \mathbf{h}_N] \equiv \Phi_L H, \tag{3}$$

and we compute Φ_L as

$$\Phi_L = [\mathbf{u}_1 \, \mathbf{u}_2 \, \cdots \mathbf{u}_N] H^{-1}, \tag{4}$$

where $H^{-1} = \frac{1}{N} H^T$. Similarly, we estimate Φ_R by projecting N patterns of the form $\mathbf{1}\mathbf{h}_k^T$.

Figure 3 depicts the calibration procedure in which we projected separable patterns on a screen and recorded sensor measurements; the sensor measurements recorded from these patterns are re-ordered to form the left and right multiplexing operators shown in (b).

A mask modulates light only by non-negative values. M-sequences are defined in terms of ± 1 values and hence cannot be directly implemented in a mask. The masks we use in our prototype cameras are constructed by computing the outer-product of two M-sequences and then setting the resulting -1 entries to 0. This produces a mask that is optically feasible but no longer mathematically separable. We can resolve this issue in post-processing, since the difference between the measurements using the theoretical ± 1 separable mask and the optically feasible $0/1$ mask is simply a constant bias term. In practice, once we acquire a sensor image, we correct it to correspond to a ± 1 separable mask (described as Y in (2)) simply by forcing the column and row sums to zero.

IV. Image Reconstruction

Given a set of $M \times M$ sensor measurements Y, our ability to invert the system (2) to recover the desired $N \times N$ image X primarily depends on the rank and the condition number of the system matrices Φ_L, Φ_R.

If both Φ_L and Φ_R are well-conditioned, then we can estimate X by solving a simple least-squares problem

$$\widehat{X}_{\text{LS}} = \arg\min_X \; \|\Phi_L X \Phi_R^T - Y\|_2^2, \tag{5}$$

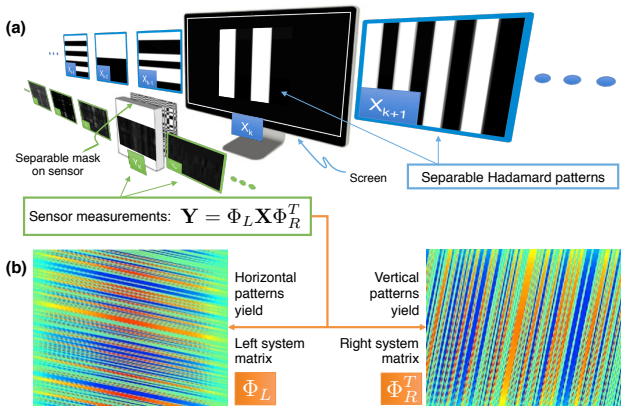


Fig. 3: Calibration for measuring the left and right multiplexing matrices Φ_L and Φ_R corresponding to a separable mask. **(a)** Display separable patterns on a screen. The patterns are orthogonal, one-dimensional Hadamard codes that are repeated along either the horizontal or vertical direction. **(b)** Estimated left and right system matrices.

which has the closed form solution: $\widehat{X}_{\text{LS}} = \Phi_L^+ Y \Phi_R^+$, where Φ_L^+ and Φ_R^+ denote the pseudoinverse of Φ_L and Φ_R, respectively. Consider the SVD of $\Phi_L = U_L \Sigma_L V_L^T$, where U_L and V_L are orthogonal matrices that contain the left and right singular vectors and Σ_L is a diagonal matrix that contains the singular values. Note that this SVD need only be computed once for each calibrated system. The pseudoinverse can then be efficiently pre-computed as $\Phi_L^+ = V_L \Sigma_L^{-1} U_L^T$.

When the matrices Φ_L, Φ_R are not well-conditioned or are under-determined (e.g., when we have fewer measurements M than the desired dimensionality of the scene N, as in compressive sensing [7]–[9]), some of the singular values are either very small or equal to zero. In these cases, the least-squares estimate $\widehat{X}_{\text{LS}}$ suffers from noise amplification. A simple approach to reduce noise amplification is to add an ℓ_2 regularization term in the least-squares problem in (5)

$$\widehat{X}_{\text{Tik}} = \arg\min_X \; \|\Phi_L X \Phi_R^T - Y\|_2^2 + \tau \|X\|_2^2, \tag{6}$$

where $\tau > 0$ is a regularization parameter. The solution of (6) can also be explicitly written using the SVD of Φ_L and Φ_R as we describe below.

The solution of (6) can be computed by setting the gradient of the objective in (6) equal to zero and simplifying the resulting equation:

$$\Phi_L^T (\Phi_L X \Phi_R^T - Y) \Phi_R + \tau X = 0$$
$$\Phi_L^T \Phi_L X \Phi_R^T \Phi_R + \tau X = \Phi_L^T Y \Phi_R.$$

Replacing Φ_L and Φ_R with their SVD decompositions yields

$$V_L \Sigma_L^2 V_L^T X V_R \Sigma_R^2 V_R^T + \tau X = V_L \Sigma_L U_L^T Y U_R \Sigma_R V_R^T.$$

Multiplying both sides of the equation with V_L^T from the left and V_R from the right yields

$$\Sigma_L^2 V_L^T X V_R \Sigma_R^2 + \tau V_L^T X V_R = \Sigma_L U_L^T Y U_R \Sigma_R.$$

Denote the diagonal entries of Σ_L^2 and Σ_R^2 using the vectors σ_L and σ_R, respectively, to simplify the equations to

$$V_L^T X V_R \odot (\sigma_L \sigma_R^T) + \tau V_L^T X V_R = \Sigma_L U_L^T Y U_R \Sigma_R$$
$$V_L^T X V_R \odot (\sigma_L \sigma_R^T + \tau \mathbf{1}\mathbf{1}^T) = \Sigma_L U_L^T Y U_R \Sigma_R$$
$$V_L^T X V_R = (\Sigma_L U_L^T Y U_R \Sigma_R)./(\sigma_L \sigma_R^T + \tau \mathbf{1}\mathbf{1}^T),$$

where $A \odot B$ and $A./B$ denote element-wise multiplication and division of matrices A and B, respectively. The solution of (6) can finally be written as

$$\widehat{X}_{\text{Tik}} = V_L[(\Sigma_L U_L^T Y U_R \Sigma_R)./(\sigma_L \sigma_R^T + \tau \mathbf{1}\mathbf{1}^T)]V_R^T. \quad (7)$$

Thus, once the SVDs of Φ_L and Φ_R are computed and stored, reconstruction of an $N \times N$ image from $M \times M$ sensor measurements involves a fixed cost of two $M \times N$ matrix multiplications, two $N \times N$ matrix multiplications, and three $N \times N$ diagonal matrix multiplications.

In many cases, exploiting the sparse or low-dimensional structure of the unknown image significantly enhances reconstruction performance. Natural images and videos exhibit a host of geometric properties, including sparse gradients and sparse coefficients in certain transform domains. Wavelet sparse models and total variation (TV) are widely used regularization methods for natural images [28], [29]. By enforcing these geometric properties, we can suppress noise amplification as well as obtain unique solutions. A pertinent example for image reconstruction is the sparse gradient model, which can be represented in the form of the following total-variation (TV) minimization problem:

$$\widehat{X}_{\text{TV}} = \arg\min_X \|\Phi_L X \Phi_R^T - Y\|^2 + \lambda \|X\|_{\text{TV}}. \quad (8)$$

The term $\|X\|_{\text{TV}}$ denotes the TV of the image X given by the sum of magnitudes of the image gradients. Given the scene X as a 2D image, i.e., $X(u, v)$, we can define $G_u = D_u X$ and $G_v = D_v X$ as the spatial gradients of the image along the horizontal and vertical directions, respectively. The total variation of the image is then defined as

$$\|X\|_{\text{TV}} = \sum_{u,v} \sqrt{G_u(u,v)^2 + G_v(u,v)^2}.$$

Minimizing the TV as in (8) produces images with sparse gradients. The optimization problem (8) is convex and can be efficiently solved using a variety of methods. Many extensions and performance analyses are possible following the recently developed theory of compressive sensing.

In addition to simplifying the calibration task, separability of the coded mask also significantly reduces the computational burden of image reconstruction. Iterative methods for solving the optimization problems described above require the repeated application of the multiplexing matrix and its transpose. Continuing our numerical example from above, for a non-separable, dense mask, both of these operations would require on the order of 10^{12} multiplications and additions for megapixel images. With a separable mask, however, the application of the forward and transpose operators requires only on the order of 2×10^9 scalar multiplications and additions—a tremendous reduction in computational complexity.

V. Experimental results

We present results on two prototypes. The first uses a Silcon-based sensor to sense in visible wavelengths and the second uses an InGaAs sensor for sensing in short-wave infrared.

A. Visible wavelength FlatCam prototype

We built this FlatCam prototype as follows.

Image sensor: We used a Sony ICX285 CCD color sensor that came inside a Point Grey Grasshopper 3 camera (model GS3-U3-14S5C-C). The sensor has 1036×1384 pixels, each 6.45μm wide, arranged in an RGB Bayer pattern. The physical size of the sensor array is approximately 6.7mm $\times$ 8.9mm.

Mask material: We used a custom-made chrome-on-quartz photomask that consists of a fused quartz plate, one side of which is covered with a pattern defined using a thin chrome film. The transparent regions of the mask transmit light, while the chrome film regions of the mask block light.

Mask pattern and resolution: We created the binary mask pattern as follows. We first generated a length-255 M-sequence consisting of ± 1 entries. The actual 255-length M-sequence is shown in Fig. 5. We repeated the M-sequence twice to create a 510-length sequence and computed the outer product with itself to create a 510×510 matrix. Since the resulting outer product consist of ± 1 entries, we replaced every -1 with a 0 to create a binary matrix that is optically feasible. An image showing the final 510×510 mask pattern is shown in Fig. 5. We printed a mask from the 510×510 binary matrix such that each element corresponds to a $\Delta = 30\mu$m square box (transparent, if 1; opaque, if 0) on the printed mask. Images of the pattern that we used for the mask and the printed mask are presented in Fig. 5. The final printed mask is a square approximately 15.3mm on a side and covers the entire sensor area. Even though the binary mask is not separable as is, we can represent the sensor image using the separable system described in (2) by subtracting the row and column mean from the sensor images (see Sec. III-D for details on calibration).

Mask placement: We opened the camera body to expose the sensor surface and placed the quartz mask on top of it using mechanical posts such that the mask touches the protective glass (hot mirror) on top of the sensor. Thus the distance between the mask and the sensor d is determined by thickness of the glass, which for this sensor is 0.5mm.

Data readout and processing: We adjusted the white balance of the sensor using Point Grey FlyCapture software and recorded images in 8-bit RGB format using suitable exposure and frame rate settings. In most of our experiments, the exposure time was fixed at 10ms, but we adjusted it according to the scene intensity to avoid excessively bright or dark sensor images. For the static scenes we averaged 20 sensor images to create a single set of measurements to be used for reconstruction.

We reconstructed 512×512 RGB images from our prototype using 512×512 RGB sensor measurements. Since the sensor has 1086×1384 pixels, we first cropped and uniformly subsampled the sensor image to create an effective 512×512 color sensor image; then we subtracted the row and column

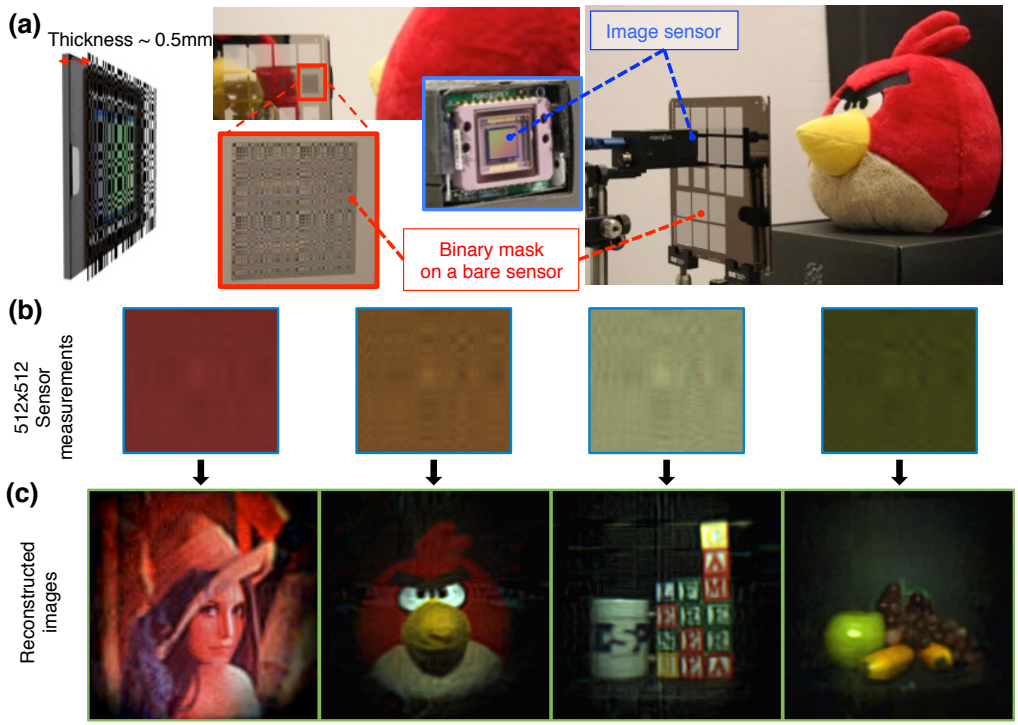


Fig. 4: Visible FlatCam prototype and results. **(a)** Prototype consists of a Sony ICX285 sensor with a separable M-sequence mask placed approximately 0.5mm from the sensor surface. **(b)** The sensor measurements are different linear combinations of the light from different points in the scene. **(c)** Reconstructed 512×512 color images by processing each color channel independently.

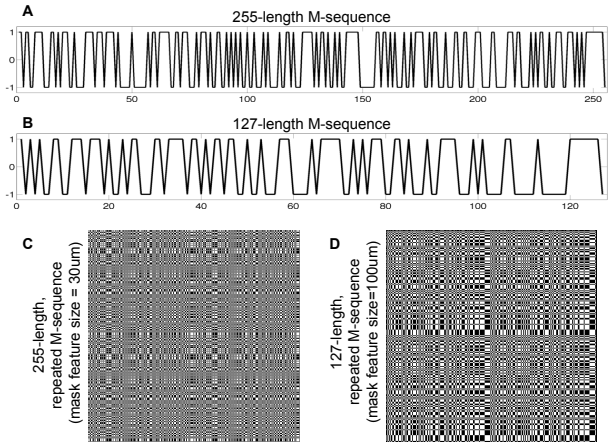


Fig. 5: Masks used in both our visible and SWIR FlatCam prototypes. M-sequences with ± 1 entries that we used to create the binary masks for **(a)** the visible camera and **(b)** the SWIR camera. Binary masks created from the M-sequences for **(c)** the visible camera and **(d)** the SWIR camera.

means from that image. The resulting image corresponds to the measurements described by (2), which we used to reconstruct the desired image X. Some example sensor images and corresponding reconstruction results are shown in Fig. 4. In these experiments, we solved an ℓ_2-regularized least-squares problem in (6), followed by BM3D denoising [30]. Solving the least-squares recovery problem for a single 512×512 image using pre-computed SVD requires a fraction of a second on a standard laptop computer.

We present a comparison of three different methods for reconstructing static scenes in Fig. 6. We used MATLAB for solving all the computational problems. For the results presented in Fig. 6, we recorded sensor measurements while displaying test images on an LCD monitor placed 28cm away from the camera and by placing various objects in front of the camera in ambient lighting.

We used three methods for reconstructing the scenes from the sensor measurements:

1) We computed and stored the SVD of Φ_L, Φ_R and solved the ℓ_2-regularized problem in (6) as described in (7). The average computation time for the reconstruction of a single 512×512 image on a standard laptop was 75ms. The results of SVD-based reconstruction are presented in Fig. 6B. The reconstructed images are slightly noisy, with details missing around the edges.
2) To reduce the noise in our SVD-estimated images, we applied BM3D denoising to each reconstructed image. The results of SVD/BM3D reconstruction are presented in Fig. 6C. The average computation time for BM3D denoising of a single image was 10s.
3) To improve our results further, we reconstructed by solving the TV minimization problem (8). While, as expected, the TV method recovers more details around the edges, the overall reconstruction quality is not appreciably very different from SVD-based reconstruction. The computation time of TV, however, increases to 75s per image.

To demonstrate the flexibility of our FlatCam design, we also captured and reconstructed dynamic scenes at typical video rates. We present selected frames[1] from two videos in

[1]Complete videos are available at `http://bit.ly/FlatCam`.

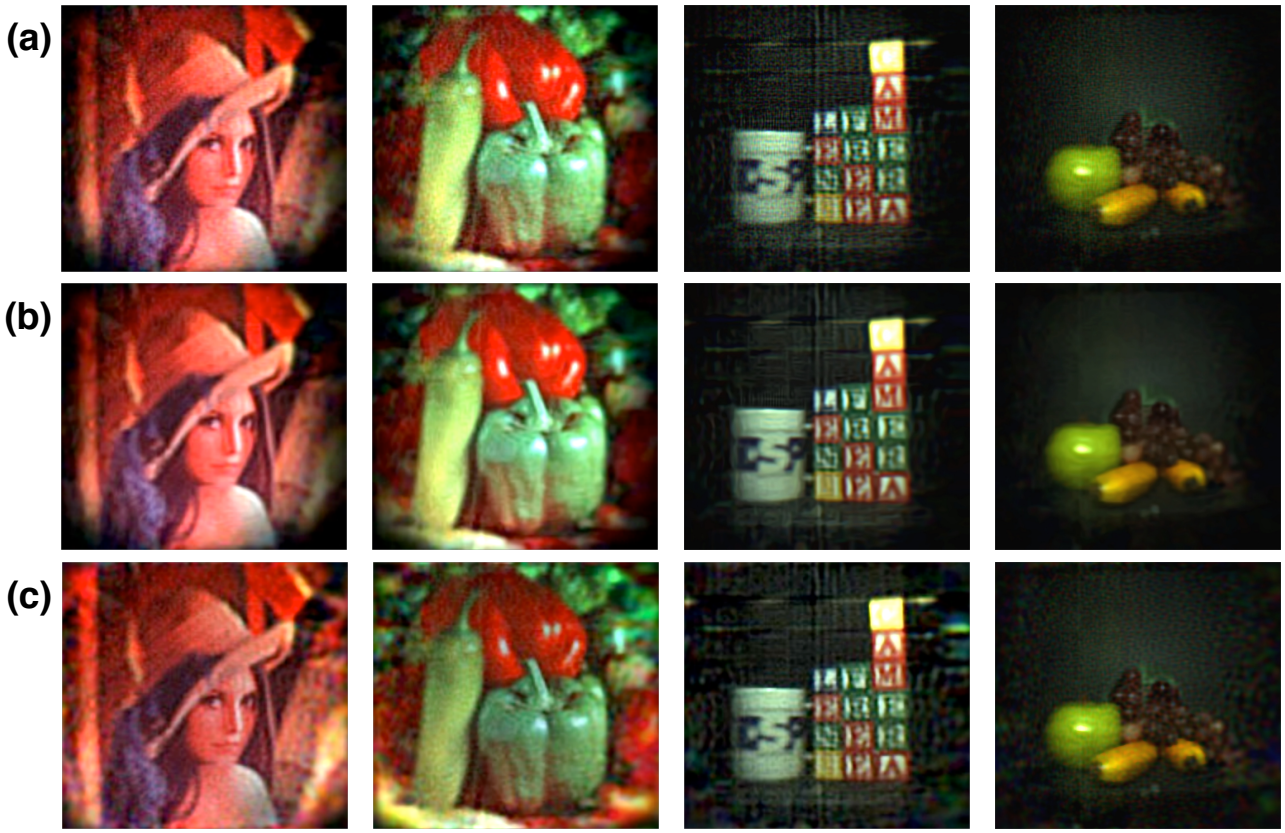


Fig. 6: Images reconstructed at 512×512 resolution using the visible FlatCam prototype and three different reconstruction methods. **(a)** SVD-based solution of (6); average computation time per image = 75ms. **(b)** SVD/BM3D reconstruction; average computation time per image = 10s. **(c)** Total variation (TV) based reconstruction; average computation time per image = 75s.

Fig. 7. The images presented in Fig. 7A are reconstructed frames from a video of a hand making counting gestures, recorded at 30 frames per second. The images presented in Fig. 7B are reconstructed frames from a video of a toy bird dipping its head in water, recorded at 10 frames per second. In both cases, we reconstructed each video frame at 512×512 pixel resolution by solving (6) using the SVD-based method described in (7), followed by BM3D denoising.

B. SWIR FlatCam prototype

This FlatCam prototype consists of a Goodrich 320KTS-1.7RT InGaAs sensor with a binary separable M-sequence mask placed at distance $d = 5$mm. The sensor-mask distance is large in this prototype because of the protective casing on top of the sensor. We used a feature size of $\Delta = 100\mu$m for the mask, which was constructed using the same photomask process as for the visible camera. The sensor has 256×300 pixels, each of size $w = 25\mu$m, but because of the large sensor-to-mask distance and mask feature size, the effective system resolution is limited. Therefore, we binned 4×4 pixels on the sensor (and cropped a square region of the sensor) to produce sensor measurements of effective size 64×64. We reconstructed images with the same 64×64 resolution; example results are shown in Fig. 8.

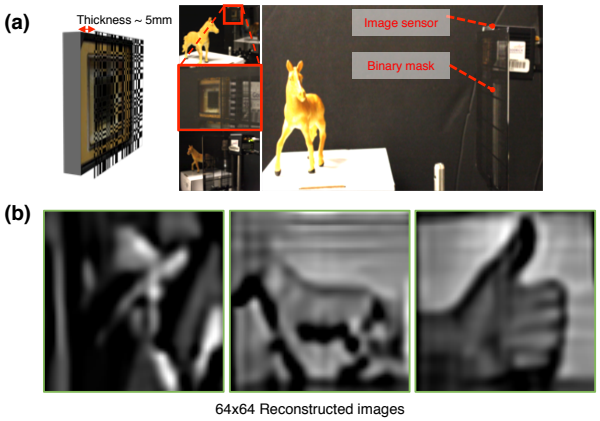


Fig. 8: Short wave infrared (SWIR) FlatCam prototype and results. **(a)** Prototype consists of a Goodrich 320KTS-1.7RT sensor with a separable M-sequence mask placed approximately 5mm from the detector surface. **(b)** Reconstructed 64×64 images.

VI. Discussions and Conclusions

The mask-based, lens-free FlatCam design proposed here can have a significant impact in an important emerging area of imaging, since high-performance, broad-spectrum cameras can be monolithically fabricated instead of requiring cumbersome post-fabrication assembly. The thin form factor and low cost

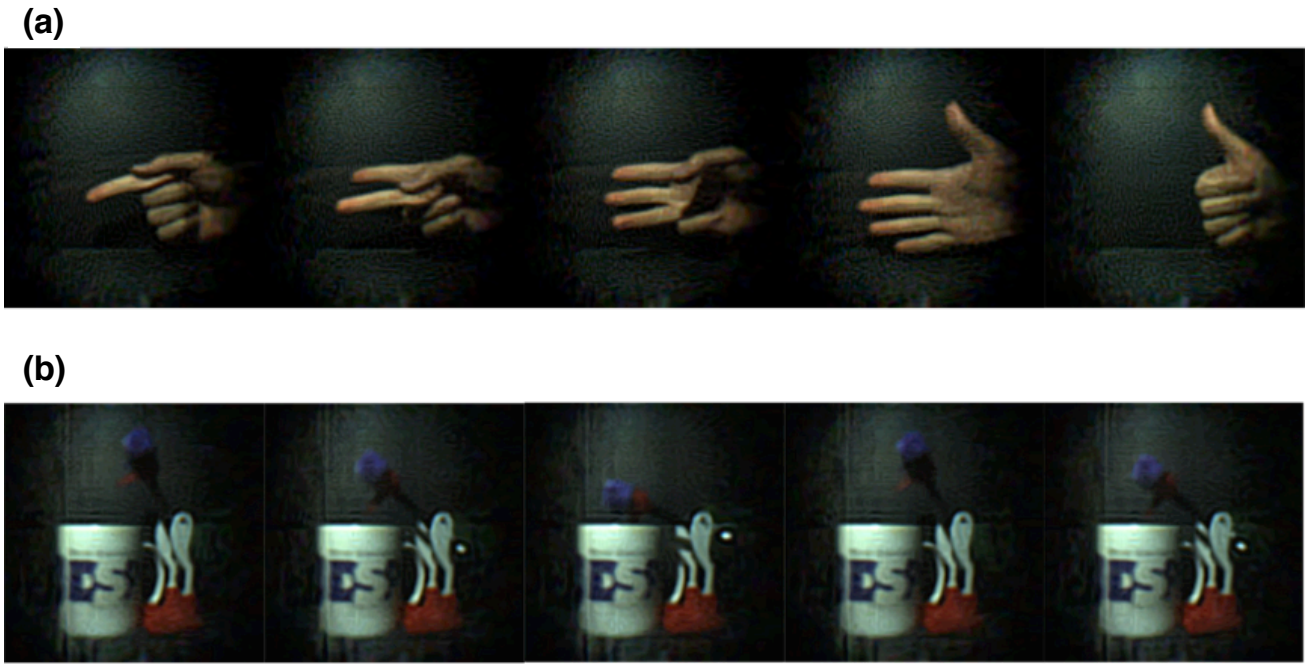


Fig. 7: Dynamic scenes captured by a FlatCam at video rates and reconstructed at 512×512 resolution. **(a)** Frames from the video of hand gestures captured at 30 frames per second. **(b)** Frames from the video of a toy bird captured at 10 frames per second.

of lens-free cameras makes them ideally suited for many applications in surveillance, large surface cameras, flexible or foldable cameras, disaster recovery, and beyond, where cameras are either disposable resources or integrated in flat or flexible surfaces and therefore have to satisfy strict thickness constraints. Emerging applications like wearable devices, internet-of-things, and in-vivo imaging could also benefit from the FlatCam approach.

A. Advantages of FlatCam

We make key changes in our FlatCam design to move away from the cube-like form-factor of traditional lens-based and coded aperture cameras while retaining their high light collection abilities. We move the coded mask extremely close to the image sensor, which results in a thin, flat camera. We use a binary mask pattern with 50% transparent features, which, when combined with the large surface area sensor, enables large light collection capabilities. We use a separable mask pattern, similar to the prior work in coded aperture imaging [15], which enables simpler calibration and reconstruction. The result is a radically different form factor from previous camera designs that can enable integration of FlatCams into large surfaces and flexible materials such as wallpaper and clothes that require thin, flat, and lightweight materials [31].

Flat form factor. The flatness of a camera system can be measured by its thickness-to-width ratio (TWR). The form factor of most cameras, including lens-based cameras, conventional coded-aperture systems, pinhole cameras, and miniature diffraction grating-based cameras, is cube-like; that is, the thickness of the device is of the same order of magnitude as the sensor width, resulting in TWR ≈ 1. Cube-like camera systems suffer from a significant limitation: if we reduce the thickness of the camera by an order of magnitude while preserving its TWR, then the area of the sensor drops by two order of magnitude. This results in a two orders of magnitude reduction in light collection ability. In contrast, FlatCams are endowed with flat form factors; by design, the thickness of the device is an order of magnitude smaller than the sensor width. Thus, for a given a thickness constraint, a FlatCam can utilize a large sensing surface for light collection. In our visible FlatCam prototype, for example, the sensor-to-mask distance is 0.5mm, while the sensor width is about 6.7mm, resulting in TWR ≈ 0.075. While on-chip lensless microscopes can also achieve such low TWRs, such systems require complete control of the illumination and the subject to be less than 1mm from the camera. We are unaware of any other far-field imaging system that has a comparable TWR of the FlatCam while providing reasonable light capture and imaging resolution.

High light collection. The light collection ability of an imaging system depends on two factors: its sensor area and the square of its numerical aperture. Conventional sensor pixels typically have an angular response of 40–60 degrees, which is referred to as the sensors chief ray angle (CRA). The total amount of light that can be sensed by a sensor is often limited by the CRA, which in turn determines the maximum allowable numerical aperture of the system. Specifically, whether we consider the best lens-based camera, or even a fully exposed sensor, the cone of light that can enter a pixel is determined by the CRA.

Consider an imaging system with a strict constraint on the device thickness T_{max}. The light collection L of such an imaging device can be described as $L \propto W^2 N_A^2$, where W denotes the width of the (square) sensor and N_A denotes

the numerical aperture. Since $W_{\max} = T_{\max}/\text{TWR}$, we have $L \propto W^2 N_A^2 \leq (N_A T_{\max}/\text{TWR})^2$. Thus, given a thickness constraint $T_{\max}$, the light collection of an imaging system is directly proportional to the square of the numerical aperture and inversely proportional to the square of its TWR. Thus, smaller TWR leads to better light collection.

The numerical aperture of our prototype FlatCams is limited by the CRA of the sensors. Moreover, half of the features in our mask are opaque and block one half of the light that would have otherwise entered the sensor. Realizing that the numerical aperture of such a FlatCam is reduced only by a factor of $\sqrt{2}$ compared to an open aperture, yet its TWR is reduced by an order of magnitude leads to the conclusion that a FlatCam collects approximately two orders of magnitude more light than a cube-like miniature camera of the same thickness.

B. Limitations of FlatCam

FlatCam is a radical departure from centuries of research and development in lens-based cameras, and as such this radical departure has its own limitations.

Achievable image/angular resolution. Our current prototypes have low spatial resolution which is attributed to two factors. First, it is well known that angular resolution of pinhole cameras and coded aperture cameras decreases when the mask is moved closer to the sensor [6]. This results in an implicit tradeoff between the achievable thickness and the achievable resolution. Second, the image recorded on the image sensor in a FlatCam is a linear combination of the scene radiance, where the multiplexing matrix is controlled by the mask pattern and distance between mask and sensor. This means that recovering the scene from sensor measurements requires demultiplexing. Noise amplification is an unfortunate outcome of any linear demultiplexing based system. While the magnitude of this noise amplification can be controlled by careful design of the mask patterns, they cannot be completely eliminated in FlatCam. In addition, the singular values of the linear system are such that the noise amplification for higher spatial frequencies is larger, which consequently limits the spatial resolution of the recovered image. We are currently working on several techniques to improve the spatial resolution of the recovered images.

Direct-view and real-time operation. In traditional lens-based cameras, the image sensed by the image sensor is the photograph of the scene. In FlatCam, a computational algorithm is required to convert the sensor measurements into a photograph of the scene. This results in a time-lag between the sensor acquisition and the image display, a time-lag that depends on processing time. Currently, our SVD-based reconstruction operates at near real-time (about 10 fps) resulting in about a 100 ms delay between capture and display. While this may be acceptable for certain applications, there are many other applications such as augmented reality and virtual reality, where such delays are unacceptable. Order of magnitude improvements in processing times are required before FlatCam becomes amenable to such applications.

ACKNOWLEDGMENTS

This work was supported by NSF grants IIS–1116718, CCF–1117939, and CCF–1527501.

REFERENCES

[1] R. Dicke, "Scatter-hole cameras for x-rays and gamma rays," *The Astrophysical Journal*, vol. 153, p. L101, 1968.

[2] E. Fenimore and T. Cannon, "Coded aperture imaging with uniformly redundant arrays," *Applied optics*, vol. 17, no. 3, pp. 337–347, 1978.

[3] T. Cannon and E. Fenimore, "Coded aperture imaging: Many holes make light work," *Optical Engineering*, vol. 19, no. 3, pp. 193–283, 1980.

[4] P. Durrant, M. Dallimore, I. Jupp, and D. Ramsden, "The application of pinhole and coded aperture imaging in the nuclear environment," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 422, no. 1, pp. 667–671, 1999.

[5] V. Dragoi, A. Filbert, S. Zhu, and G. Mittendorfer, "Cmos wafer bonding for back-side illuminated image sensors fabrication," in *2010 11th International Conference on Electronic Packaging Technology & High Density Packaging*, 2010, pp. 27–30.

[6] D. J. Brady, *Optical imaging and spectroscopy*. John Wiley & Sons, 2009.

[7] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] R. G. Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, vol. 24, no. 4, 2007.

[10] R. F. Marcia and R. M. Willett, "Compressive coded aperture superresolution image reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 833–836.

[11] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied optics*, vol. 47, no. 10, pp. B44–B51, 2008.

[12] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *Optics express*, vol. 21, no. 9, pp. 10 526–10 545, 2013.

[13] A. Zomet and S. K. Nayar, "Lensless imaging with a controllable aperture," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 339–346.

[14] G. Huang, H. Jiang, K. Matthews, and P. Wilford, "Lensless imaging by compressive sensing," in *20th IEEE International Conference on Image Processing*, 2013, pp. 2101–2105.

[15] M. J. DeWeert and B. P. Farm, "Lensless coded-aperture imaging with separable doubly-toeplitz masks," *Optical Engineering*, vol. 54, no. 2, pp. 023 102–023 102, 2015.

[16] J. Tanida, T. Kumagai, K. Yamada, S. Miyatake, K. Ishida, T. Morimoto, N. Kondou, D. Miyazaki, and Y. Ichioka, "Thin observation module by bound optics (TOMBO): concept and experimental verification," *Applied optics*, vol. 40, no. 11, pp. 1806–1813, 2001.

[17] M. Shankar, R. Willett, N. Pitsianis, T. Schulz, R. Gibbons, R. Te Kolste, J. Carriere, C. Chen, D. Prather, and D. Brady, "Thin infrared imaging systems through multichannel sampling," *Applied optics*, vol. 47, no. 10, pp. B1–B10, 2008.

[18] A. Brückner, J. Duparré, R. Leitel, P. Dannberg, A. Bräuer, and A. Tünnermann, "Thin wafer-level camera lenses inspired by insect compound eyes," *Optics Express*, vol. 18, no. 24, pp. 24 379–24 394, 2010.

[19] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An ultra-thin high performance monolithic camera array," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 166, 2013.

[20] E. J. Tremblay, R. A. Stack, R. L. Morrison, and J. E. Ford, "Ultrathin cameras using annular folded optics," *Applied optics*, vol. 46, no. 4, pp. 463–471, 2007.

[21] A. Wang, P. Gill, and A. Molnar, "Angle sensitive pixels in cmos for lensless 3d imaging," in *IEEE Custom Integrated Circuits Conference*, 2009, pp. 371–374.

[22] P. R. Gill, C. Lee, D.-G. Lee, A. Wang, and A. Molnar, "A microscale camera using direct fourier-domain scene capture," *Optics letters*, vol. 36, no. 15, pp. 2949–2951, 2011.

[23] P. R. Gill and D. G. Stork, "Lensless ultra-miniature imagers using odd-symmetry spiral phase gratings," in *Computational Optical Sensing and Imaging*. Optical Society of America, 2013, pp. CW4C–3.

[24] D. Stork and P. Gill, "Lensless ultra-miniature cmos computational imagers and sensors," in *International Conference on Sensor Technologies and Applications*, 2013, pp. 186–190.

[25] A. Greenbaum, W. Luo, T.-W. Su, Z. Göröcs, L. Xue, S. O. Isikman, A. F. Coskun, O. Mudanyali, and A. Ozcan, "Imaging without lenses: Achievements and remaining challenges of wide-field on-chip microscopy," *Nature methods*, vol. 9, no. 9, pp. 889–895, 2012.

[26] A. Greenbaum, Y. Zhang, A. Feizi, P.-L. Chung, W. Luo, S. R. Kandukuri, and A. Ozcan, "Wide-field computational imaging of pathology slides using lens-free on-chip microscopy," *Science translational medicine*, vol. 6, no. 267, pp. 267ra175–267ra175, 2014.

[27] S. W. Golomb, *Shift register sequences*. Aegean Park Press, 1982.

[28] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[29] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[30] 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, 2007.

[31] F. Koppens, T. Mueller, P. Avouris, A. Ferrari, M. Vitiello, and M. Polini, "Photodetectors based on graphene, other two-dimensional materials and hybrid systems," *Nature nanotechnology*, vol. 9, no. 10, pp. 780–793, 2014.