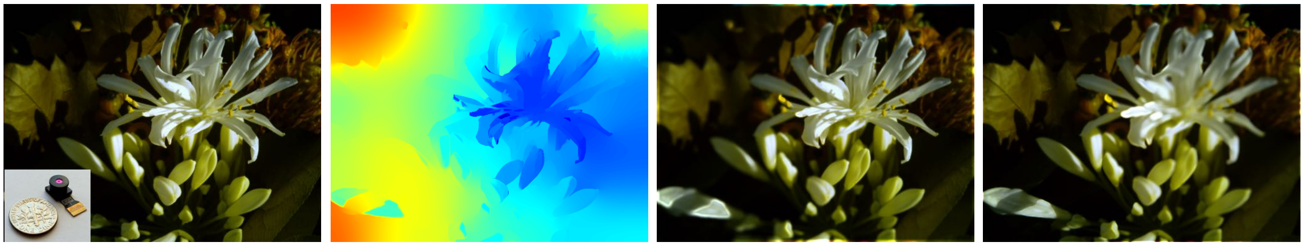
Dual Aperture Photography: Image and Depth from a Mobile Camera

Manuel Martinello[1] Andrew Wajs[1] Shuxue Quan[1] Hank Lee[1] Chien Lim[1]
Taekun Woo[1] Wonho Lee[2] Sang-Sik Kim[3] David Lee[1]

[1]Dual Aperture International [2]Silicon File Technologies, Inc. [3]SK Hynix, Inc.



(a) Extended DOF image (with DA camera) (b) Depth map (c) Refocused Close (d) Refocused Far

Figure 1. **Outputs of a mobile DA camera.** All-in-focus image (a) and depth map (b) of a scene with flowers using the DA camera for mobile devices displayed in the inset at the bottom left of (a). The image can be refocused after the capture as shown in (c) and (d).

Abstract

Conventional cameras capture images with limited depth of field and no depth information. Camera systems have been proposed that enable additional depth information to be captured with the image. These systems reduce the resolution of the captured image or result in reduced sensitivity of the lens. We demonstrate a camera that is able to capture extended depth of field images together with depth information at each single frame while requiring minimal impact on the physical design of the camera or its performance. In this paper we show results with a camera for mobile devices, but this technology (named dual aperture *to recall the major change in the camera model) can be applied with even greater effect in larger form factor cameras.*

1. Introduction

The vast majority of cameras that are shipped today are embedded in mobile devices. These cameras have constraints on physical size, mechanical parts and processing capacity. Several solutions have been proposed to measure depth and extend the Depth of Field (DOF), but they usually involve relatively large devices (typically based on a DSLR form factor), reduce light sensitivity and spatial resolution, or require the capture of multiple frames. We have designed and developed a novel computational camera able to extend the DOF of an image and measure depth; The most important, it can do so by capturing a single image and it is compact enough to be incorporated into a mobile device. The sensor of the proposed system captures infrared image data in addition to red, green, and blue image components. The camera has a second narrower aperture for the infrared part of the light spectrum: from here its name, *Dual Aperture* (DA) camera. This results in the infrared channel having an image with a larger DOF than the visible channels. This difference is used to estimate the relative blur between the IR and RGB channels, which provides a measure of distance of an object from the lens (Figure 1(b)); it can also be used to remove the blur from the visible image components, as shown in Figure 1(a).

DA camera enables applications such as image refocussing, generation of 3D image pairs, 3D reconstruction, and gesture tracking to be embedded within a mobile device. In addition the proposed camera can perform these functions under difficult conditions such as bright sunlight. The device is versatile and might open a new research area in computer vision, enabling algorithms that make use of the combination of data from both the visible and invisible parts of the spectrum.

2. Related Work

In the past years several techniques have been proposed to 1) extend the DOF of the image and 2) recover the depth map of the scene. When trying to achieve these goals by using only a single image, most of the work overlaps since to extend the DOF the blur shape is needed, and blur is where the depth information is usually encoded.

978-1-4799-8667-5/15/$31.00 ©2015 IEEE

Some depth invariant blurs have been designed to perform deconvolution without depth estimation. In wavefront coding this is obtained by a cubical optical element [11] at the expense of an increased dynamic range of the incoherent system. Other works study the use of a logarithmic asphere [7] or a focus sweep, where the focus is modified during the image integration by moving the image detector of a camera [22] or the specimen under a microscope [17]. More recently, an extended DOF image has been restored from a camera without the use of moving parts, either by using a diffuser [9] or by exploiting axial chromatic aberrations [16, 8]. In [16] the sharper area of a color channel is successfully transferred locally to the other channels [15], but the image details are limited by the quality of the best focused channel, which can be blurred. In [8] the authors propose to maximize the axial chromatic aberration and invert the defocus blur without explicitly estimating depth. A sharper image can be obtained by deconvolving the luminance channel of the captured image with a depth-invariant PSF, although this leaves the chrominance channels with residual blur and chromatic aberration. Instead, in our work we first estimate the depth map by recovering the blur size at each pixel of the image: we can then deblur all the channels when forming the all-in-focus image.

One of the most successful approaches is the capture of 4D radiance into a 2D sensor within a single photograph. This is achieved by placing an array of lenses in front of a conventional camera [13], or between the main lens and the sensor in a plenoptic camera [2, 24]. Depth information is extracted to allow to refocus after the capture. Both approaches trade spatial resolution for the ability to solve angular differences. This can be partially overcome by superresolution [4, 12], but its complexity is very high and significant artefacts reduce the practical resolution of the camera. A cheaper approach is given by coded aperture photography, where a mask is placed on the lens aperture to make the blur shape easier to identify. Depth and all-in-focus image can be extracted from a single shot [18, 19, 28, 31] or from a pair of images [30]. The main disadvantage is that the mask blocks some light going through the lens, thus decreasing the signal to noise ratio (SNR) of the captured image.

Some systems designed exclusively for depth measurements make use of additional illumination to achieve their goal, especially in textureless regions. A known pattern of visible [21] or infrared light [29] is projected into the scene; objects distort the pattern based on their distance from the camera. In another popular active system the time-of-flight camera emits a the near-infrared wave and estimates depth by measuring the phase delay of the same wave when reflected by the scene [25]. These systems require an active light source and a dedicated IR camera and they do not work in sunlight conditions where the ambient IR radiation overwhelms the IR from the light source.

A very interesting approach is presented in [14], where a multi-aperture camera uses mirrors to divide the image into four sections: each section captures the same scene with a different aperture (hence different DOF). The different images are then combined into a single one, whose focus can be manipulate by analyzing the information contained in the four captured images. The use of mirrors has a significant impact on the form factor of the camera, and the effective sensor resolution is reduced by a factor of four.

Our work has similarity to [6], where a color filter is mounted inside a DSLR camera to reduce the aperture of the green channel only. Depth and deblurred image are extracted by comparing images with different DOF. This approach presents some limitations when applied to a mobile camera: 1) to achieve a meaningful depth resolution, the green aperture should be further reduced, which would yield to a consistent reduction of the total incoming light since the green pixels form 50% of a conventional sensor; 2) The reduced intensity of the green component affects the dynamic range of the green channel for any illumination.

In this paper we propose a novel technology to overcome most of these limitations, in particular we show that:

1. DA camera is more light efficient and can be mounted and used successfully in mobile devices (Section 3);
2. the dynamic range (mainly of the R channel) is reduced only for lights with high IR (Section 4.4);
3. the proposed depth estimation approach is more accurate and more robust to noise then [6] (Section 5);
4. near-IR data of the scene is available together with the usual RGB image.

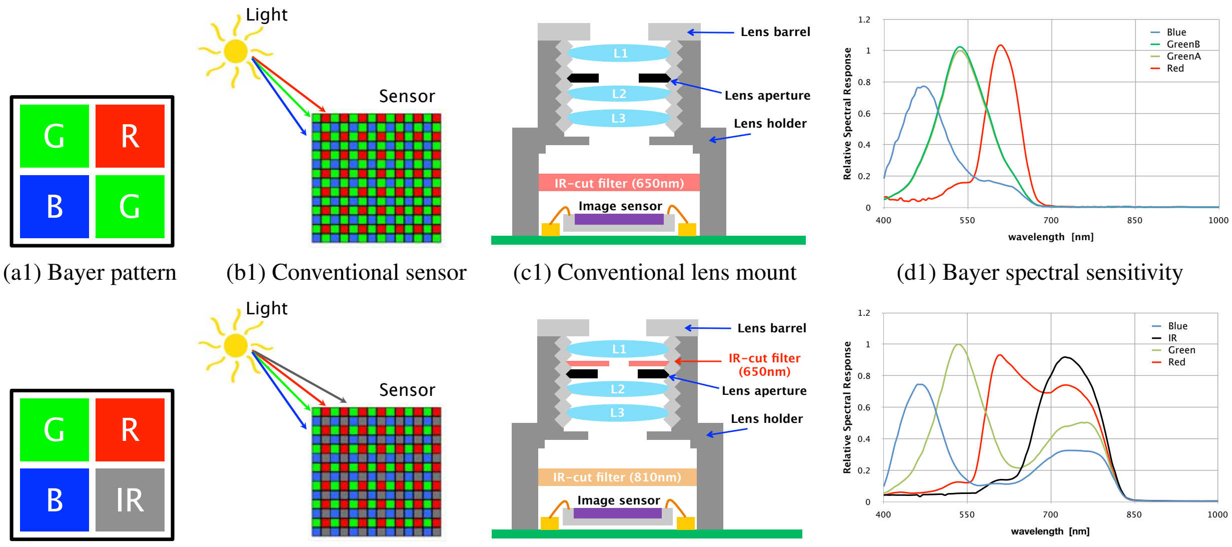
3. Dual Aperture Camera Model

The DA camera can be obtained by applying two alterations to any conventional camera: 1) enable the sensor to read IR data in addition to visible light, and 2) create two apertures in the lens, one for visible light and one for near-infrared light. Only light from the visible parts of the spectrum is passed through the wider aperture while light from visible and near-IR is passed through the narrower aperture. This results in the sensor being able to capture an image where the IR channel has a larger DOF than the other three channels. By comparing these two types of images depth and all-in-focus image can be extracted (Section 5).

These main differences with a conventional camera are described in details in Section 3.1 and Section 3.2 and illustrated in Figure 2. The effect of the IR aperture size on the image quality is also addressed in Section 3.3.

3.1. RGB-IR Sensor

A conventional digital camera utilizes a sensor that is made up of light sensitive pixels. Normally these pixels are coated with dyes composed of 3 colors - red, green, and



(a1) Bayer pattern (b1) Conventional sensor (c1) Conventional lens mount (d1) Bayer spectral sensitivity

(a2) DA pattern (b2) DA sensor (c2) DA lens mount (d2) DA spectral sensitivity

Figure 2. **Conventional camera (top row) vs. DA camera (bottom row).** (a-b) Difference in the pattern and the camera sensor; (c) Difference in the lens mount; (d) Response of each channel for difference wavelengths using a conventional Bayer sensor (top) and the proposed RGB-IR sensor (bottom).

blue. Each pixel only captures light from one of the three colors. Currently one of the most common pattern used to laid out the colored pixels in the sensors is the 2 × 2 *Bayer pattern* (shown in Figure 2(a1-b1)), which comprises a red, a blue, and two green pixels. The color image is reconstructed through a process, called *demosaicing*, whereby the missing colors for each pixel are reconstituted using the information from adjacent pixels.

In the proposed camera the sensor is modified such that the Bayer block structure remains, with one of the green pixels being replaced by an IR pixel (see Figure 2(a2-b2)). This enables the new sensor to capture IR information as well as RGB information. Ideally the infrared frequencies should only be stored in the IR pixels, but in practise there is a very strong *cross-talk* between RGB and IR, as illustrated in Figure 2(d2). This problem is addressed in the image pipeline by an IR removal task, described in Section 4.4.

3.2. Dual Aperture Lens

In a conventional camera the infrared signal is entirely blocked from reaching the sensor by placing an IR-cut filter (∼650nm) on top of it, as illustrated in Figure 2(c1).

In the DA camera we replace this filter with one that blocks only frequencies greater than 810nm, allowing near-infrared (from 650nm to 810nm) to reach the sensor together with visible light. At the same time we mount at the center of the lens a glass disk coated with IR blocking material (∼650nm) with a hole that constitutes the IR aperture (Figure 2(c2)). The glass is opaque to the near-IR wavelengths and has the effect of creating a narrower aperture for IR light only; it allows instead the visible light to pass resulting in the RGB aperture of the lens being unaffected by the presence of the glass disk. The small aperture causes the IR data to have a low SNR. This problem will be solved by a nonlocal-mean denoising approach in Section 4.1.

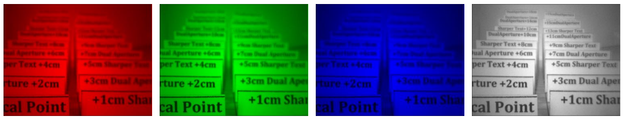
3.3. IR Aperture Selection

Ideally the IR aperture should be: 1) as large as possible to limit the loss of incoming light and therefore the noise in the image; 2) as small as possible to reduce the IR cross-talk into RGB and have an IR channel with a wider DOF than the RGB image. Clearly both criteria cannot be completely satisfied: We search for the optimal balance between them.

When the radius of the lens aperture is reduced by a factor of α, with $\alpha < 1$, from its original size A, the light reaching the sensor is reduced by a factor of $(1 - \alpha^2)$. Suppose there are N_a smaller apertures in the lens, which affect different color channels, the total reduction of the light reaching the sensor is equal to the sum of the effects caused by each aperture:

$$L_r = \sum_{k=1}^{N_a} \frac{n_k * (1 - \alpha_k^2)}{4}. \quad (1)$$

where n_k corresponds to the number of pixels (in the sensor pattern) affected by the aperture modification and the denominator indicates the total number of pixels in the pattern (shown in Figure 2(a1-a2)).



(a) Red Channel (b) Green channel (c) Blue Channel (d) IR channel ($\times 5$)

Figure 3. **RGB vs. IR channels.** IR channel has a larger DOF since the light rays are going through a smaller aperture.

Using equation (1) we can quantitatively compare our solution with the two most similar camera systems recently proposed in literature. Chakrabarti and Zickler [6] reduce only the aperture for the green channel of a factor $\alpha_1 = 0.59$; since they use a Bayer sensor this alteration affects two pixels in the sensor pattern ($n_1 = 2$) and causes a reduction of light efficiency of $L_r = 32.6\%$. Instead, Bando *et al.* [3] suggest to split the camera lens into three ($N_a = 3$) smaller color-filtered apertures (red, blue, and green) and $\alpha_k = 0.5$ for each of them, with the green aperture affecting two pixels in the Bayer pattern; compared with its original dimension, this device has a total light loss of $L_r = 75\%$.

Experimentally we found that in our camera a filter with $\alpha_1 = 0.46$ for the IR aperture provides the best balance between our above defined criterions: The total light efficiency is reduced by only $L_r = 20\%$ and the difference in the DOF between the IR and RGB channels is large enough to provide accurate depth information. Figure 3 shows the four channels of an image captured by a DA camera: notice the deeper DOF of the IR compared to the visible channels.

4. Image Reconstruction

The creation of a conventional RGB image is performed by using steps similar to those used in a conventional camera, as shown in Figure 4. Some of the processing is modified to compensate for the changes made to the camera. Red blocks in the flow diagram represent tasks which have been modified or designed specifically for the DA camera. Green blocks instead represent tasks identical to the ones already used in a conventional camera. (e.g., the demosaicing process takes as input an image with a 3-channel Bayer pattern and generates a full-resolution RGB image).

In this section we describe the tasks of image reconstruction which have been modified to suit a DA camera.

4.1. Noise reduction

When capturing an image there is some noise η added to the original radiance $\boldsymbol{I}$, especially to the IR channel which, in general, has a lower intensity since it goes through a narrower aperture. Therefore, assuming that η is a white Gaussian noise of standard deviation σ, our goal is to recover the original image $\boldsymbol{I}$ from the recorded noisy image $\boldsymbol{I}_\eta$:

$$\boldsymbol{I}_\eta = \boldsymbol{I} + \eta_\sigma \ . \tag{2}$$

For every pixel $\mathbf{p} \in \boldsymbol{I}_\eta$ we find all pixels that resemble $\mathbf{p}$ by using a nonlocal-means filter [5]. The resemblance is evaluated by comparing patches $\boldsymbol{g}$ around the selected pixel:

$$\boldsymbol{I}(\mathbf{p}) = \frac{1}{Z(\mathbf{p})} \sum_{\mathbf{q} \in N_\mathbf{p}} \mathcal{W}(\mathbf{p}, \mathbf{q}) \, \boldsymbol{I}_\eta(\mathbf{q}) \ , \tag{3}$$

where $Z(\mathbf{p}) \doteq \sum_\mathbf{q} \mathcal{W}(\mathbf{p}, \mathbf{q})$ is the normalization factor, $N_\mathbf{p}$ is the neighbourhood of the pixels $\mathbf{p}$, and the weights $\mathcal{W}$ are given by

$$\mathcal{W}(\mathbf{p}, \mathbf{q}) = e^{-\frac{G_\rho * |\boldsymbol{g}_\eta(\mathbf{p}) - \boldsymbol{g}_\eta(\mathbf{q})|^2(0)}{\tau}} ; \tag{4}$$

τ represents the bandwidth of the filter, G_ρ is the Gaussian kernel with standard deviation ρ and

$$G_\rho * |\boldsymbol{g}_\eta(\mathbf{p}) - \boldsymbol{g}_\eta(\mathbf{q})|^2(0) = \int_{\mathbb{R}^2} G_\rho(\boldsymbol{t}) |\boldsymbol{g}_\eta(\mathbf{p}+\boldsymbol{t}) - \boldsymbol{g}_\eta(\mathbf{q}+\boldsymbol{t})|^2 d\boldsymbol{t} \ . \tag{5}$$

In other words, equation (3) means that the value $\boldsymbol{I}(\mathbf{p})$ is replaced by a weighted average of $\boldsymbol{I}_\eta(\mathbf{q})$. The weights are significant only if a Gaussian window around $\mathbf{q}$ is similar to the corresponding window around $\mathbf{p}$. Assuming that self-similarity extends throughout the signal, the neighborhood $N_\mathbf{p}$ is ideally taken to be the entire image, so the averaging process is fully nonlocal [5, 27]. However, for computational purposes we restrict the search of similar patches (of size 3×3) in a larger search window of size 9×9. The algorithm can efficiently be implemented using the parallel architecture of current devices [10].

4.2. IR interpolation

The infrared channel records an extended DOF image of the scene in grayscale, as shown in Figure 3. To match this sharp information to the RGB domain we extract the missing IR pixels by using a bicubic interpolation (to preserve the sharpness of the edges) and obtain a full resolution IR image, as illustrated in Figure 5(a). This image can now be easily overlay to all the other channels.

4.3. Green Pixel Replacement

We want to recreate the Bayer pattern shown in Figure 2(a1) so that common image processing tasks can be used. Thus every pixel of the IR channel has to be replaced with a green one. We name these new pixels G_2 to distinguish them from the green pixels G originally captured. The replacement of the pixel $\mathbf{p}$ is achieved by interpolating the intensities of the four nearest G pixels (diagonal directions) together with the information of the interpolated IR channel. Since it has been shown that the color-difference space yields to better quality than the original color space [1, 26], we interpolate the space $K = G - IR_G$. The pixels at the location G_2 are then obtained as

$$G_2 = K + IR. \tag{6}$$

This procedure is illustrated in Figure 5(b).

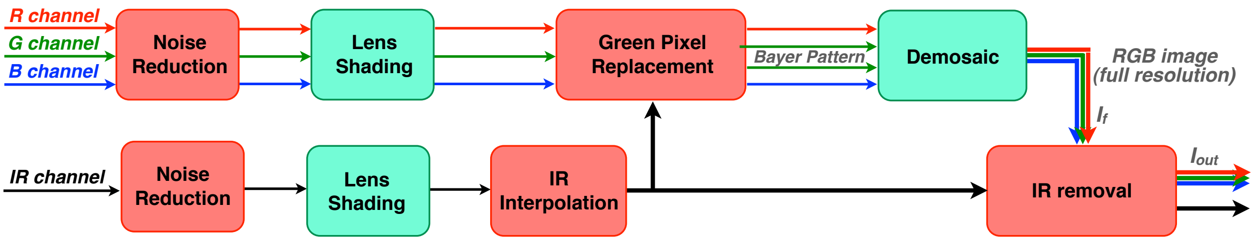
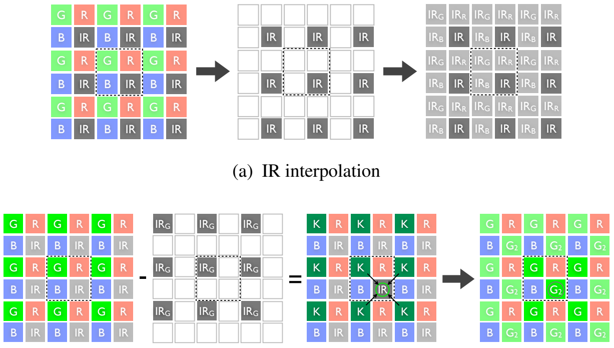


Figure 4. **Image reconstruction pipeline.** The green blocks represent tasks which are unchanged from a conventional camera, while the red block are tasks that have been created or modified ad-hoc for the DA camera. Thicker lines indicate full resolution images.



(a) IR interpolation

(b) Green pixel replacement

Figure 5. **IR interpolation and green pixel replacement.**

4.4. IR Removal

The RGB-IR sensor of the proposed camera presents a *cross-talk* mainly from the IR to the RGB channels (Figure 2(d2)). If neglected, this can distort the colors of the image when the illumination contains near IR components. Hence, we want to remove from the RGB channels the contribution of the near IR wavelengths (650 − 810nm). The visible channels are affected differently but all of them contain a percentage of the image stored in the IR channel (with the red channel having almost the full IR image added to its signal).

The final RGB image $\boldsymbol{I}_{out} = [\boldsymbol{I}_{out}^R\ \boldsymbol{I}_{out}^G\ \boldsymbol{I}_{out}^B]^T$ is then obtained by subtracting at each visible channel i of the full resolution image $\boldsymbol{I}_f = [\boldsymbol{I}_f^R\ \boldsymbol{I}_f^G\ \boldsymbol{I}_f^B]^T$ different proportions α_i of the IR channel:

$$\boldsymbol{I}_{out}^i = \boldsymbol{I}_f^i - \alpha_i \boldsymbol{I}_f^{IR}, \qquad i \in \{R,\ G,\ B\}\,. \qquad (7)$$

To compute the coefficients α_i we analyse the spectral sensitivity of the DA sensor in Figure 2(d2). For a given wavelength λ, the proportion of the IR cross-talk into the i-th channel is represented by the ratio between the values of the spectral response f_i and the IR spectral response f_{IR}. Hence, when considering the cross-talk happening in the whole range of the near IR ($\lambda \in [650, 810]$nm), we measure the ratio of the areas below the curves f_i and f_{IR}, and the weights α_i are so obtained

$$\alpha_i = \frac{\int_{650}^{810} f_i(\lambda)d\lambda}{\int_{650}^{810} f_{IR}(\lambda)d\lambda}\,. \qquad (8)$$

Notice that to use α_i from equation (8) in equation (7) we assume that the contribution from the wavelengths in the visible range ($\lambda < 650$nm) to the IR channel can be ignore.

After the IR removal, the image $\boldsymbol{I}_{out}$ can be process using conventional algorithms for color correction and white balance. An example of the effect of the IR subtraction is reported in the supplementary material.

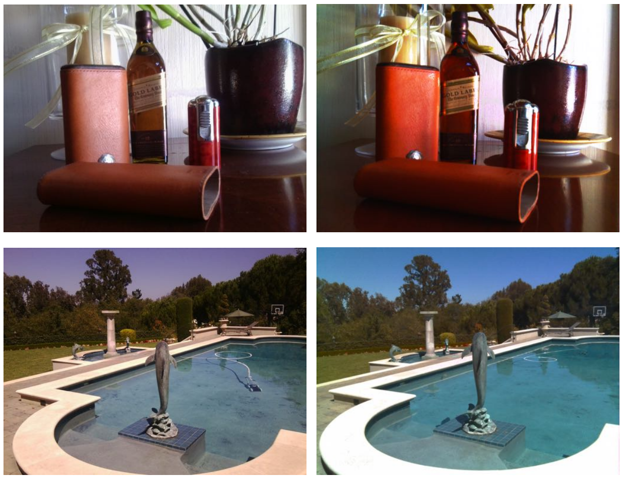
4.5. Discussion on Image Quality

To show the image quality of a DA camera, and particularly the effect of substituting a green pixel with an IR in the Bayer pattern, we have captured the same scenes with a DA camera and with the Samsung smartphone GT-S5230, which uses the same sensor (but with the original Bayer pattern). The white balance has been set to auto for both cameras. The photos are showed side-by-side in Figure 6. The exposure time of the image capture is not impacted by the camera modification. Although the images captured with the DA camera show a slightly higher noise level, the final colors reproduce very well the original natural scene.

A possible future approach to further improve the quality of the images through the image pipeline, could be to combine demosaicing, green pixel replacement, and IR interpolation tasks, in-line with the method described in [23]. However, for the sake of efficiency in this work we have kept these tasks separated in order to use some algorithms already optimized for mobile cameras.

4.5.1 Dynamic range

When the source of illumination has a low emission of near-IR, the IR channel is noisy but its intensity is very low. Therefore there is almost no impact in the dynamic range of the visible channels. Instead, for light sources with a high amount of near-IR (e.g., sunlight or incandescent), the



(a) Samsung GT-S5230 (Bayer pattern) (b) Dual Aperture camera

Figure 6. **Image quality comparison.** The same scenes have been captured by (a) a mobile camera with a Bayer sensor and (b) a DA camera using the same sensor but with modified pattern.

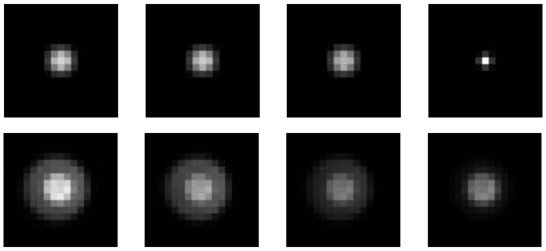
SNR of the IR image is higher but the IR cross-talk into the visible channels reduces the dynamic range, mainly of the red channel.

In the supplementary material we report a side-by-side color quality comparison between a conventional camera (Bayer pattern) and a DA camera under the standard illuminance D65 (similar to day sunlight), which has been chosen because of its wide spectrum and strong level of near-infrared. The small difference in color distances when compared with the traditional RGB camera might be explained by the reduced dynamic range in the red channel.

5. Depth and All-In-Fucus Image Estimation

When an object is placed at the focus distance, the object appears sharp in all the four channels. If an object is placed away from the focus distance, it appears defocused and the amount of defocus is proportional to its distance from the focal plane. Assuming the object is located at depth $\boldsymbol{d}_k$, the captured image will be defocused with the blur $\boldsymbol{h}^{IR}_{\boldsymbol{d}_k}$ in the IR component and with the blur $\boldsymbol{h}^{i}_{\boldsymbol{d}_k}$ in the channel i, with $i = \{R, G, B\}$. Because of the difference in the size of the lens aperture, the blur $\boldsymbol{h}^{IR}_{\boldsymbol{d}_k}$ is always equal (when object is in-focus) or smaller (when object is out-of-focus) than the blur $\boldsymbol{h}^{i}_{\boldsymbol{d}_k}$. Hence, the set of kernels $\boldsymbol{h}^{IR}_{\boldsymbol{d}_k}$ and $\boldsymbol{h}^{i}_{\boldsymbol{d}_k}$ uniquely identifies the depth $\boldsymbol{d}_k$.

For simplicity, in this section we assume that $\boldsymbol{d}(\mathbf{p})$ indicates the depth value at the pixel $\mathbf{p}$. It actually represents the index k of the set of kernels and therefore the relative distance from the focal plane; once the focus setting of the camera is known, the absolute depth value can be obtained from $\boldsymbol{d}(\mathbf{p})$, as described in the supplementary material.



(a) R channel (b) G channel (c) B channel (d) IR channel

Figure 7. **Blur kernels of a DA camera.** Examples of real PSFs of a DA camera for an illumination with a high amount of near-IR (incandescent). Top row: small blur size (blur level# 7); Bottom: large blur size (blur level# 19).

5.1. PSF Calibration

In a DA camera there is a strong cross-talk between the IR and the RGB channels (Figure 2(d2)), which can be corrected at the end of the image pipeline as described in Section 4.4. However, when estimating depth we use the unprocessed images and this cross-talk affects significantly the shape of the blur kernels differently at each channel and depending on the amount of near-IR emitted by the light source (see Figure 7).

To have an accurate estimation of the PSFs we do not use synthetic kernels but capture the PSF for different light conditions using the following procedure, which takes into account the new aperture weighting, cross-talks, and different IR levels: 1) Place a black cloth with a very small hole in front of a source of the selected illumination and take a picture; 2) The captured image gives us the shape and the intensity distribution of the PSF for all four channels; 3) We repeat the capturing changing the focus of the camera or the distance camera - light source. When performing depth estimation, we use the auto white balance to determine the type of illumination and choose the matching set of PSFs.

5.2. Depth Estimation

Given the observed (unprocessed) 4-channel image $\boldsymbol{I} = \left[\boldsymbol{I}^R\ \boldsymbol{I}^G\ \boldsymbol{I}^B\ \boldsymbol{I}^{IR}\right]^T$, we extract depth information using a 3-step approach. For every $\boldsymbol{d}_k$ that we consider, we: 1) deblur $\boldsymbol{I}^{IR}$ with the small blur $\boldsymbol{h}^{IR}_{\boldsymbol{d}_k}$, 2) blur the sharp IR image $\hat{\boldsymbol{f}}^{IR}_k$ with the kernel $\boldsymbol{h}^{i}_{\boldsymbol{d}_k}$, and then 3) compute the match between the synthetically blurred image and the captured channel $\boldsymbol{I}^i$ at every pixel. Pixels belonging to objects at depth $\boldsymbol{d}_{\hat{k}}$ should have the best match when $k = \hat{k}$.

The first step is achieved by solving a least square optimization problem

$$\hat{\boldsymbol{f}}^{IR}_k = \underset{\boldsymbol{f}^{IR}_k}{\operatorname{argmin}} \left[\left| \left(\boldsymbol{h}^{IR}_{d_k} * \boldsymbol{f}^{IR}_k\right) - \boldsymbol{I}^{IR} \right|^2 + \beta\, E_p(\boldsymbol{f}^{IR}_k) \right] , \tag{9}$$

which searches for the sharp texture $\boldsymbol{f}_k^{IR}$ that minimizes the reconstruction error and the texture prior E_p, preferring $\boldsymbol{f}_k^{IR}$ to be as smooth as possible:

$$E_p(\boldsymbol{f}_k^{IR}) = \left|\nabla_x \boldsymbol{f}_k^{IR}\right|^2 + \left|\nabla_y \boldsymbol{f}_k^{IR}\right|^2 \ . \tag{10}$$

In equation (9) we assume Gaussian distribution so that it can be quickly minimized in closed form [18].

Once we have the sharp image $\hat{\boldsymbol{f}}_k^{IR}$ we blur it with the correspondent larger kernel $\boldsymbol{h}_{\boldsymbol{d}_k}^i$ of channel i. In our implementation we found the good depth estimation results can be obtained only considering the green channel: $i = \{G\}$.

The third step is a challenging task since the pixel intensities of two channels can be very different. We choose to compute the Normalized Cross Correlation (NCC) of the sum of the gradients at each pixel $\mathbf{p}$:

$$E_m(\mathbf{p}, k) = NCC\left(\nabla_{xy}((\boldsymbol{h}_{\boldsymbol{d}_k}^G * \hat{\boldsymbol{f}}_k^{IR})_{\mathbf{p}})\ ,\ \nabla_{xy}(\boldsymbol{I}_{\mathbf{p}}^G)\right) \tag{11}$$

where $\nabla_{xy}(u) = \nabla_x(u) + \nabla_y(u)$, and the sub-indexed $u_{\mathbf{p}}$ indicates a patch extracted from the image u, centred in $\mathbf{p}$ and of the same size of the kernels $\delta \times \delta$.

The depth value at each pixel is obtained by finding the blurs that maximizes the match in equation (11)

$$\boldsymbol{d}(\mathbf{p}) = \underset{k}{\text{argmax}}\ E_m(\mathbf{p}, k) \ . \tag{12}$$

The resulting raw depth map is noisy as showed in the example in Figure 8(c). Similarly to the technique used for the noise reduction in equation (3), we combine depth information of pixels that belong to similar regions

$$\boldsymbol{d}_s(\mathbf{p}) = \frac{1}{Z(\mathbf{p})} \sum_{\mathbf{q} \in N_{\mathbf{p}}} \mathcal{W}(\mathbf{p}, \mathbf{q})\ \boldsymbol{d}(\mathbf{q}) \tag{13}$$

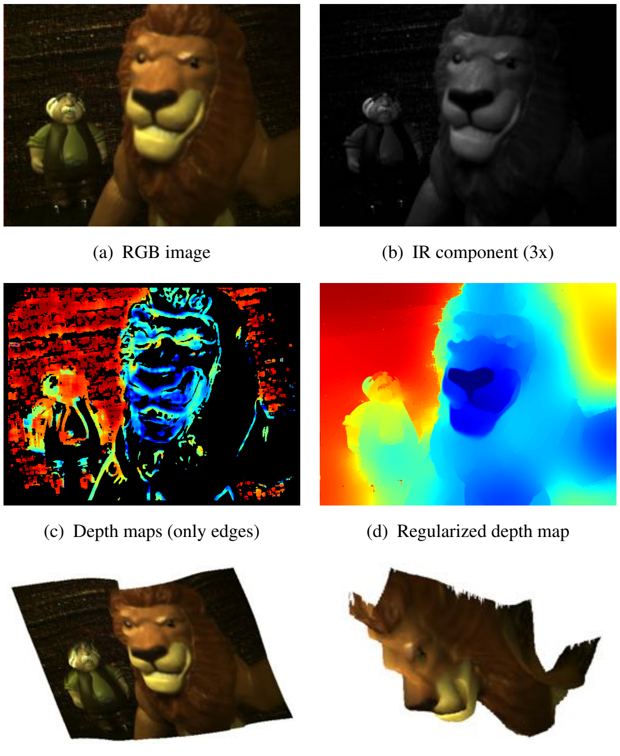
where the weights $\mathcal{W}$ are given by

$$\mathcal{W}(\mathbf{p}, \mathbf{q}) = e^{-\frac{G_\sigma * |\boldsymbol{I}_{\mathbf{p}}^{IR} - \boldsymbol{I}_{\mathbf{q}}^{IR}|^2(0)}{\tau}} \tag{14}$$

using $\boldsymbol{I}^{IR}$ in order to have sharper boundaries. In other words, depth is extracted by taking the weighted average of the raw depth values at the neighbouring pixels only if they share the same texture. This yields to a much smoother depth map, such as the one in Figure 8(d). Using this depth map realistic 3D reconstructions of the captured scene can be successfully obtained (Figure 8(e) and Figure 8(f)).

5.3. Limitations of the Depth Estimation

This approach is based on the assumption that edges in the visible domain have corresponding edges in the IR image. However, there are textures that do not respect this assumption. If the difference in intensities at the edge is comparable to the level of noise, the edge cannot be identified. We solve this problem by considering only edges whose (absolute) gradient is greater than a given threshold. For the rare case when an edge in the visible domain does not have a corresponding edge in the IR channel, we can still estimate an approximate depth value by applying the same procedure in equation (9) - (12) but substituting the IR image $\boldsymbol{I}^{IR}$ (and the relative kernels $\boldsymbol{h}_{d_k}^{IR}$) with the blue or red component. This is possible because the PSFs are different for each channel, due to the different IR cross-talk.



(a) RGB image

(b) IR component (3x)

(c) Depth maps (only edges)

(d) Regularized depth map

(e) 3D reconstruction

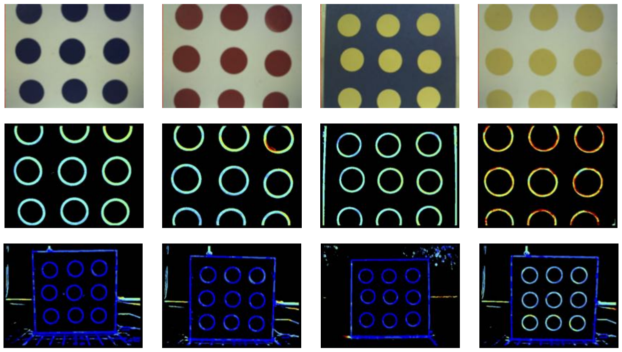
(f) 3D reconstruction

Figure 8. **Depth estimation procedure.** (a-b) Input images (unprocessed) used to extract depth information; (c) the depth is extracted at the edges using equation (12); (d) final depth map estimated from equation (13).

Due to the small difference in blur, the depth estimation for these points are not as accurate as when using the IR channel, as showed in Figure 9, where the depth values at the edges of the yellow disks (last column) have been obtained by using the red channels instead of the IR. However, we found that in general the regularization term helps recovering the correct value using the neighbouring pixels, as showed in Figure 8(d).

5.4. All-in-focus Image

Similarly to the method used for deblurring the IR channel, described in equation (9), for each considered blur scale



(a) (b) (c) (d)

Figure 9. **Limitation of our depth estimation method.** Under D55 illumination (low IR), we focus the DA camera at 12cm and capture 2 images of a flat paper with color disks, placed at 5cm (centre row) and 10cm (bottom row) from the camera. The IR image is very noisy (due to the low level of IR) and the edges of the yellow disks cannot be distinguish from the white background. Red channel is used instead of IR for the estimating the depth of the disks, yielding to incorrect raw values.

k we search for the texture $\boldsymbol{f}_k^i$ that minimizes the functional

$$\hat{\boldsymbol{f}}_k^i = \underset{\boldsymbol{f}_k^i}{\operatorname{argmin}} \left[\left| \left(\boldsymbol{h}_{d_k}^i * \boldsymbol{f}_k^i \right) - \boldsymbol{I}^i \right|^2 + \beta\, E_p(\boldsymbol{f}_k^i) \right] \; ; \quad (15)$$

As in [18], to achieve a higher deblurred image quality we assume sparse derivatives prior for the RGB texture

$$E_p(\boldsymbol{f}_k^i) = \sum_{\mathbf{p}} \left(\left| \nabla_x \boldsymbol{f}_k^i(\mathbf{p}) \right|^{0.8} + \left| \nabla_y \boldsymbol{f}_k^i(\mathbf{p}) \right|^{0.8} \right) \; . \quad (16)$$

From our experiments, values of the raw depth map are sufficient to produce a visually plausible deblurred image, although they do not correspond to the right blur scale. Hence we can pick each pixel independently from the $\boldsymbol{f}_k^i$ with the smallest reconstruction error, and construct the all-in-focus image using the raw values from equation (12)

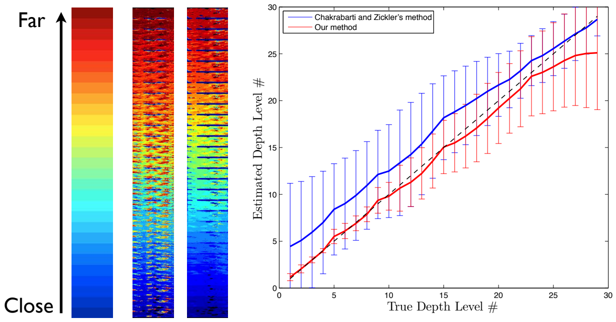
$$\hat{\boldsymbol{I}}^i(\mathbf{p}) = \hat{\boldsymbol{f}}_{d(\mathbf{p})}^i(\mathbf{p}) \; . \quad (17)$$

6. Experimental Results

We now demonstrate the effectiveness of our approach on both synthetic and real data. We also analyse the performance of our depth estimation algorithm and show that, for a DA camera, it outperforms the closest prior work.

6.1. Performance Analysis

We choose to perform a similar performance analysis than [20], where a real texture is blurred with all the 29 kernels considered in this work. Each blurred image is then attached to the others, forming a tall image with increasing blur size going from bottom to top. We then estimate the depth map (or blur map) from this image using both our approach and the closest method in literature from [6]. In this approach the authors compute a kernel $\bar{\boldsymbol{h}}_{\boldsymbol{d}_k}$ which compensates for the variation between the different per-channel blurs; In our implementation of their work, we use $\bar{\boldsymbol{h}}_{\boldsymbol{d}_k}$ as the compensation between the blurs $\boldsymbol{h}_{\boldsymbol{d}_k}^{IR}$ and $\boldsymbol{h}_{\boldsymbol{d}_k}^{i}$.



(a) Ground-Truth vs [6] vs our (b) Graph (mean ± 3std)

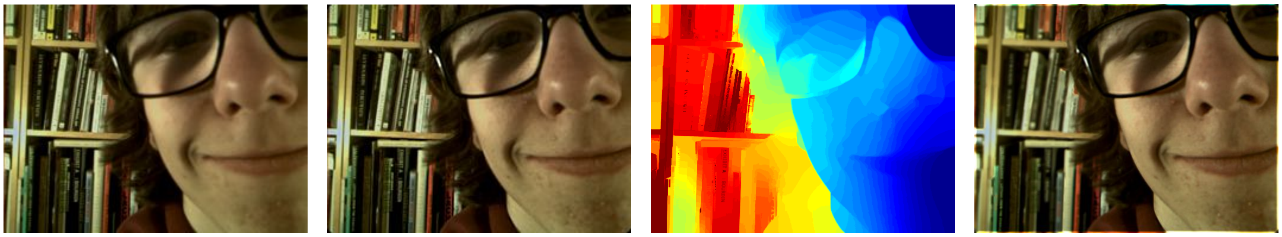
Figure 10. **Depth estimation performance.** Results on real texture and comparison with prior work [6].

Ground-truth is showed in Figure 10(a) (left), together with the results recovered by prior work (centre) and our approach (right). Both mean and standard deviation of the estimated blur scale are also shown in Figure 10(b) using an error-bar with the algorithms performances (solid line) over the ideal characteristic curve (diagonal dashed line). A more detailed analysis with different textures and types of noise is described in the supplementary material. Compared to [6], our algorithm gives more accurate estimation (particularly when the blur size is small) and it is more robust to noise, which is essential when dealing with a low intensity IR image.

6.2. Results on Real Data

All the photos in this work have been captured using the DA camera for mobile phones shown in the inset of Figure 1(a). The 3.6mm lens has an aperture of $f/2.8$ and the sensor is based upon $\frac{1}{4}''$ 3M pixel sensor that has been already deployed in mobile devices. The pixel size is $1.7\mu m$. The sample images have been captured under different lighting conditions: outdoor in the shade (Figure 1), indoor artificial energy saving 3500k lights (Figure 11 and Figure 8), and natural outdoor light and incandescent 2800k light (see examples in the supplementary material). As showed in these examples, it is possible to determine depth and generate good quality images with a DA camera under diverse lighting conditions.

The capture of depth in conjunction with the image enables several photographic applications including refocussing of the image after it has been captured (as shown in Figure 1 and Figure 11(d)) and 3D image generation. The latter task can be shown as a scene that goes *into* the screen,



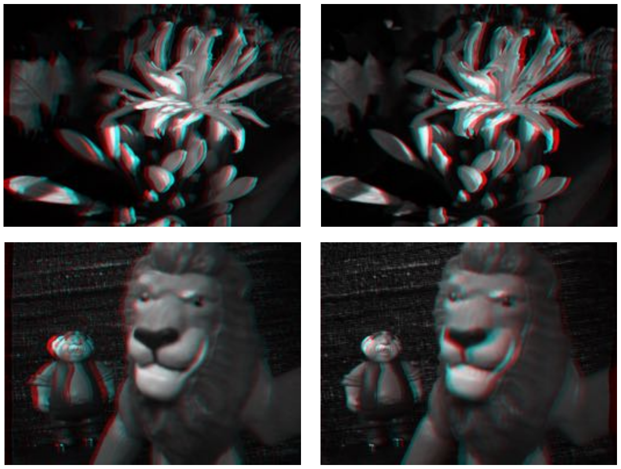
(a) Captured RGB image

(b) All-in-focus image

(c) Depth map

(d) Refocused image

Figure 11. **Adam dataset.** Picture captured indoor under artificial energy saving light. Focus plane is set at the background. Red color on the depth map indicates objects away from the focal plane, therefore closer to the camera.



(a) Scene *goes into* the screen

(b) Scene *comes out* from the screen

Figure 12. **3D images *into* and *out* of the screen.** Top: images generated from results in Figure 1; Bottom: images obtained from the dataset in Figure 8. Images to be watched with red-cyan glasses.

as in Figure 12(a) or *out* of the screen as Figure 12(b).

The depth range showed in these examples varies from 5 to 40 cm (Figure 8) and from 20 to 80cm (Figure 11) from the camera. With a larger camera it is possible to extend the depth range and its accuracy, since it depends proportionally on the camera focal length, the physical size of the lens aperture, and the number of pixel in the sensor.

6.3. Discussion on Depth Resolution

The range and accuracy of the depth measurement is determined by the properties of the optical module. The one used for these results is an existing off-the-shelf module designed for mobile applications. As with all mobile lenses the short focal length of the lens means that the hyper focal distance is relatively short and this limits the range and accuracy of the depth measurement. Figure 13 displays the depth resolution graph for the lens used in this experiments (blue curve) in the range between 5cm and 2m from the camera. The same data are illustrated on the right as depth levels: the depth of objects placed within the same color segment cannot be distinguished. Our technology can also be easily applied to different camera devices if a better depth accuracy is needed: e.g., by using the DA technology with an $8mm$ lens f/2.8 the depth resolution improves drastically (green curve in the graph).

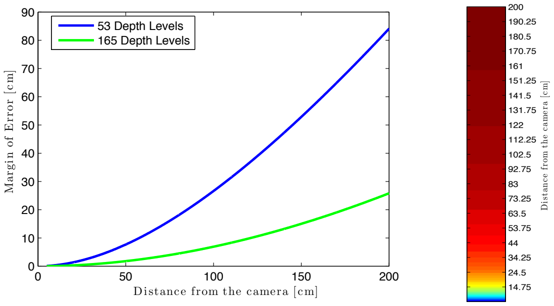


Figure 13. **Depth resolution in the range [5-200cm].** Depth resolution for a DA camera with two different lenses: 3.6mm with f/2.8 (blue curve) and 8mm with f/2.8 (green). The first setting has been used for all the experiments in this paper.

6.3.1 Comparison with Stereo Cameras

The main goal of this work is to demonstrate that depth information can be extracted, together with the RGB image, using a single mobile camera without the use of additional power. This allows one to have further information about the captured scene, even if limited to a small range of distances. DA camera cannot compete with stereo regarding depth resolution (the baseline for DA camera is the diameter of the aperture while for stereo is the distance between the two cameras), but can instead *complement* stereo. In stereo, the larger the distance between the two cameras, the better the depth resolution which can be achieved. However both cameras need to see the subject: if the subject is close to the device it might be visible only by one camera or create large occluded areas which would prevent from extracting depth information using a conventional stereo approach. Having a stereo module composed by 2 DA cameras might solve this problem, since a depth map is available from each camera and this can compensate for the lack of overlap in the views.

7. Conclusions and Future Work

In this paper we have introduced a novel camera, the Dual Aperture camera, capable of capturing an all-in-focus image and depth of the scene in a single shot. We have shown that it is feasible to modify the existing image signal processing chain to handle the dual aperture image data and that this can be performed within a mobile device. This does not result in a loss of spatial resolution or in the reduced efficiency of the lens. We have presented and implemented an algorithm that uses the differences in focus between the near-IR and RGB components to measure the depth of an object in the image. The results presented in this paper have been captured using an existing camera, originally designed for mobile devices, which has been converted into a DA camera. This demonstrates that the proposed system can be implemented in very small cameras and is extremely well suited to mobile applications.

While the proposed design has shown good performance even with real data, it remains an open question if the shape of the IR aperture can be improved to preserve more high frequencies and at the same time increase the SNR of the IR channel. Finally, the fact that the camera captures near-IR data together with RGB data can be a benefit for many other applications in the field of computer vision.

Acknowledgements

We would like to thank Silicon File Technologies and SK Hynix for their assistance in developing the camera, building the prototypes and enabling these experiments to be performed. We would also like to thank the reviewers for their constructive feedback that helped to improve this work.

References

[1] J. Adams. Design of practical color filter array interpolation algorithms for digital cameras. *Proceeding of SPIE*, 3028:117–125, 1997.

[2] T. Adelson and J. Wang. Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:99–106, 1992.

[3] Y. Bando, B.-Y. Chen, and T. Nishita. Extracting depth and matte using a color-filtered aperture. *ACM Trans. Graph.*, 27(5), Dec 2008.

[4] T. Bishop and P. Favaro. Light field camera: Extended depth of field, aliasing, and superresolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):972–986, May 2012.

[5] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. *IEEE Conference on Computer Vision and Patter Recognition*, 2005.

[6] A. Chakrabarti and T. Zickler. Depth and deblurring from a spectrally-varying depth-of-field. *IEEE European Conference on Computer Vision*, 7576:648–666, 2012.

[7] W. Chi and N. George. Computational imaging with the logarithmic asphere: theory. *Journal of the Optical Society of America*, 21(6), Jun 2003.

[8] O. Cossairt and S. K. Nayar. Spectral focal sweep: Extended depth of field from chromatic aberrations. *IEEE International Conference on Computational Photography*, Mar 2010.

[9] O. Cossairt, C. Zhou, and S. K. Nayar. Diffusion coding photography for extended depth of field. *ACM Trans. Graph.*, Aug 2010.

[10] J. Darbon, A. Cunha, T. Chan, S. Osher, and G. Jensen. Fast nonlocal filtering applied to electron cryomicroscopy. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1331–1334, May 2008.

[11] E. R. Dowski and T. W. Cathey. Extended depth of field through wave-front coding. *Applied Optics*, 34:1859–1866, 1995.

[12] T. Georgiev, G. Chunev, and A. Lumsdaine. Superresolution with the focused plenoptic camera. *SPIE Electronic Imaging*, Jan 2011.

[13] T. Georgiev, K. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intawala. Spatio-angular resolution tradeoffs in integral photography. *Eurographics Workshop on Rendering*, pages 263–272, 2006.

[14] P. Green, W. Sun, W. Matusik, and F. Durand. Multi-aperture photography. *ACM Trans. Graph.*, 26(3):68, 2007.

[15] F. Guichard. Advances in camera phone picture quality. *Photonics Spectra*, 41(11):50–51, Nov 2007.

[16] F. Guichard, H. P. Nguyen, R. Tessieres, M. Pyanet, I. Tarchouna, and F. Cao. Extended depth-of-field using sharpness transport across color channels. *SPIE*, 7250, Jan 2009.

[17] G. Hausler. A method to increase the depth of focus by two step image processing. *Optics Communications*, pages 38–42, 1972.

[18] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph.*, 26(3):70, Aug 2007.

[19] M. Martinello. *Coded Aperture Imaging*. PhD thesis, Heriot-Watt University, 2012.

[20] M. Martinello and P. Favaro. Single image blind deconvolution with higher-order texture statistics. In *Video Processing and Computational Video*, volume LNCS7082, pages 124–151. Springer-Verlag, 2011.

[21] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar. Active refocusing of images and videos. *ACM Trans. Graph.*, Aug 2007.

[22] H. Nagahara and S. Kuthirummal. Flexible depth of field photography. *European Conference on Computer Vision*, 2008.

[23] S. G. Narasimha and S. K. Nayar. Enhancing resolution along multiple imaging dimensions using assorted pixels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):518–530, Apr 2005.

[24] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford University CS, Apr 2005.

[25] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. *IEEE International Conference on Computer Vision*, 3:488–491, 2000.

[26] S. Pei and I. Tam. Effective color interpolation in ccd color filter array using signal correlation. *IEEE International Conference on Image Processing*, 3:488–491, 2000.

[27] B. Tracey. Nonlocal means denoising of ecg signals. *IEEE Transactions on Biomedical Engineering*, 59(9), Sep 2012.

[28] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69, Aug 2007.

[29] Z. Zalevsky, A. Shpunt, A. Maizels, and J. Garcia. Method and system for object reconstruction. *Patent WO2007043036A1*, 2007.

[30] C. Zhou, S. Lin, and S. K. Nayar. Coded aperture pairs for depth from defocus. *International Conference on Computer Vision*, Oct 2009.

[31] C. Zhou and S. K. Nayar. What are good apertures for defocus deblurring? *IEEE International Conference on Computational Photography*, Apr 2009.