A 3D Multi-Aperture Image Sensor Architecture

Keith Fife, Abbas El Gamal and H.-S. Philip Wong
Department of Electrical Engineering, Stanford University, Stanford, CA 94305-4055

***Abstract*— An image sensor comprising an array of apertures each with its own local integrated optics and pixel array is presented. A lens focuses the image above the sensor creating overlapping fields of view between apertures. Multiple perspectives of the image in the focal plane facilitate the synthesis of a 3D image at a higher spatial resolution than the aperture count. Depth resolution is shown to continue to improve with pixel scaling below the diffraction limit. Preliminary circuit implementation is described.**

I. Introduction

In applications such as robotics, biometrics, security and surveillance, there is a need to simultaneously extract both a 2D image and a 3D depth map of the scene. In recent years, several 3D imaging systems implementing a variety of techniques such as stereo-vision, motion parallax, depth-from-focus, and light detection and ranging (LIDAR) have been reported. In particular, multi-camera stereo vision systems infer depth using parallax from multiple perspectives [1], while time-of-flight sensors compute depth by measuring the delay between an emitted light pulse (e.g., from a defocused laser) and its incoming reflection [2]. These systems are relatively expensive, consume high power, and require complex camera calibration. Moreover, methods using active illumination, although highly accurate, employ large pixels which results in low spatial resolution for a given format.

In this paper, we describe an architecture for a single-chip multi-aperture image sensor that is capable of simultaneous, high resolution capture of a 2D image and 3D depth map of the scene. The sensor does not require an active illumination source. Furthermore, complex camera calibration is not required as the critical system dimensions are controlled by lithography. Consequently, it is well suited for low cost, miniaturized vision systems. Depth is inferred in a manner similar to stereo vision systems except that the correspondence problem is solved through the use of multiple, localized images of the focal plane.

In the following sections, we describe the architecture and operation of our image sensor, discuss how spatial resolution depends on various system parameters, and provide preliminary circuit implementation. A detailed analysis of sensor performance including the effect of nonidealities as well as characterization of test structures will be reported in future publications.

II. Architecture

The image sensor comprises an $m \times n$ aperture array, each with its own local optics and a $k \times k$ pixel array and readout circuit (see Fig. 1). The local optics are implemented in the dielectric stack of the integrated circuit using refractive microlenses or diffractive gratings patterned in the metal layers. The creation of independent apertures and localized pixels allows for aggressive pixel scaling, which is key to achieving high depth resolution as discussed later.

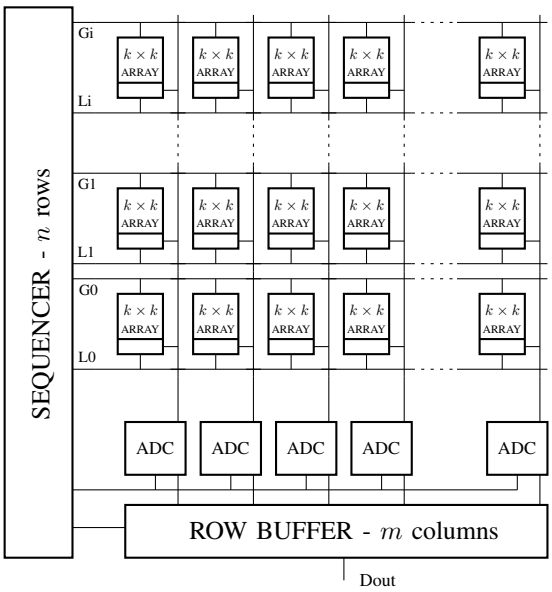
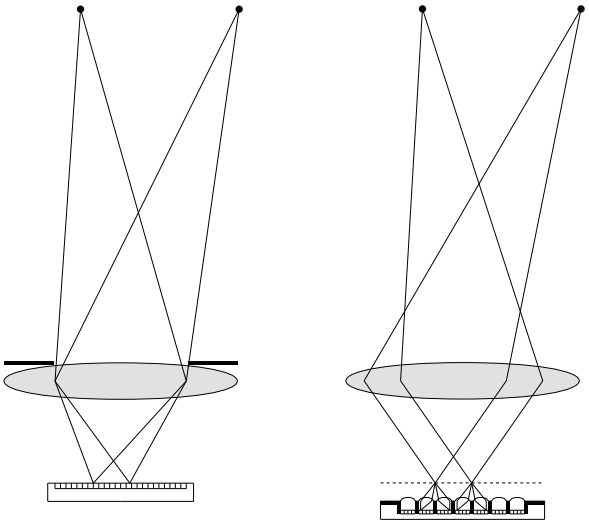


Fig. 1. Block diagram of integrated sensor.

Unlike a conventional imaging system where the lens focuses the image directly onto the image sensor (see Fig. 2(a)), the image is now focused *above* the sensor plane and re-imaged by the local optics to form partially overlapping images of the scene (see Fig. 2(b)). The captured images are combined to form the 2D and 3D representations of the scene. Note that the objective lens in our system has no aperture from the perspective of the aperture array. This allows for a relatively complete description of the wavefront in the focal plane. The amount of depth information that can be extracted depends on the total area of the objective lens that is scanned by the aperture array.

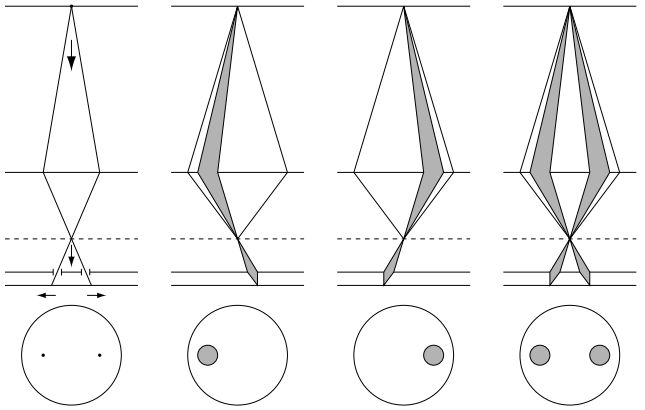
While our system is similar in structure to the plenoptic system described in [3], which employs a separate microlens array on top of an image sensor, there is a key difference between the way these two systems operate. In the plentopic system, the objective lens is focused onto the microlens array and the microlens array is focused onto the system aperture. Each microlens spreads out the incident rays to the pixels behind it, which provides information about their direction as well as intensity. While the spatial resolution of this system



(a) Traditional optical configuration.

(b) Multi-aperture optical configuration.

Fig. 2. Traditional vs multi-aperture configuration.



(a) Chief rays for a pair of apertures

(b) Left virtual objective aperture

(c) Right virtual objective aperture

(d) Virtual stereo view

Fig. 3. Virtual aperture views.

is only equal to that of the microlens array, the information about the directions of the rays can be used for a number of applications such as range finding and perspective. Note that this system contains only one aperture (that of the objective lens), which is imaged by each of the microlenses. The useful size of each pixel in this configuration is limited by microlens aberrations and fundamentally by diffraction. In contrast, our system captures less information about the wavefront but achieves higher spatial resolution than that of the aperture array. Depth is extracted by sampling the same points in the focal plane from multiple perspectives. Our image sensor architecture is also similar to that of the compound-eye [4] whose purpose is to realize a compact, thin camera with a total resolution exceeding that of an individual aperture. The spatial resolution of this scheme is largely dependent on object distance while our sensor confines the imaging to a tight region behind the objective lens to enable both high 2D and 3D spatial resolution.

III. 2D and 3D Image Extraction

Depth information is obtained from the disparity between apertures. Fig. 3(a) depicts the chief ray traces for an object as it is imaged from the apertures behind the objective lens. The circle below the diagram shows the location at which the chief ray pierces the objective lens. As the object moves back and forth, the object in the focal plane (above the sensor) moves back and forth with some attenuation in magnitude governed by the lens law. The movement of the object in the secondary images formed by the local optics is lateral. Therefore, the amount of lateral displacement between multiple apertures corresponds to the depth of the object. With several apertures accurately placed with respect to each other, the correspondence between them becomes quite reliable. The marginal ray traces for the same point as seen from the two different apertures are shown in Fig. 3(b)-3(d). The circle below each diagram shows the area of the objective lens that is used by each aperture. As can be seen, a virtual stereo pair is projected up to the plane of the objective lens. The characteristics of the apertures remain constant across the array without spatial compensation, especially as the objective lens maintains telecentricity.

At nominal object distance, only a small number of apertures sample any given point in the object plane. As the object moves to further distances, more apertures capture its information. Therefore, both the position of objects within each aperture and the total number of apertures imaging the same point are indicators of depth. Since the redundancy between apertures is localized across the focal plane, spatial resolution continues to scale by adding more apertures. To increase depth resolution, as we shall see, pixel size is scaled down even below the diffraction limit. While it is difficult to scale pixels to this level in a large, uniform array, the fact that the pixels in our sensor are grouped into smaller disjoint arrays facilitates such aggressive pixel scaling.

In Subsection III-A we quantify the depth of field for our system and in Subsection III-B, we establish the relationship between object distance and the displacement at the sensor surface. In Subsection III-C, we discuss the dependency of the available 2D and 3D spatial resolution on pixel size and local magnification. We assume an ideal, diffraction limited optical system and ideal sensor characteristics. Of course, such nonidealities should be considered when computing the real spatial resolution limits.

A. Depth of Field

To evaluate the depth of field, consider the diagram in Fig. 4. Considering the parameters defined in the figure, define the distance $E = B + C$ and the magnification factors $M = B/A$ and $N = D/C$. Because we fix the distance E for a given object range, the other variables B, C, D, M, and N are all driven by the object distance A. Given a nominal object distance A_0, we denote the other parameters by B_0, C_0, D_0,

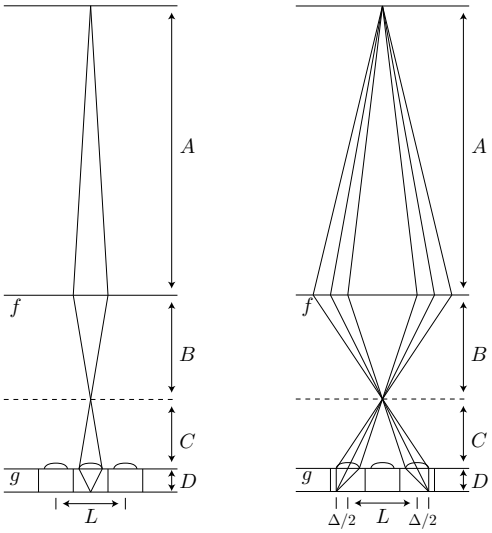


Fig. 4. Diagrams for computing depth.

M_0, and N_0. As A varies, the distance E can be adjusted to achieve the desired local magnification N_0 for the secondary image focused at D_0. This is similar to adjusting the focus in a conventional camera. The distance D_0 is approximately equal to the dielectric stack height of the fabrication process, or the nominal distance to the secondary focal plane from the local optics. Thus, given the stack height D_0, the focal length g is set during fabrication to meet the desired N_0 value. To illustrate the results, we assume that $f = 10$mm, $A_0 = 1$m, $D_0 = 10\mu$m, and $g = 8\mu$m. These parameters yield a nominal magnification factor of $N_0 = 1/4$. This value is chosen to achieve the desired amount of overlap between aperture views as detailed in later calculations.

We solve for D as a function of A by fixing the parameters to meet the nominal magnification factor N_0. To characterize the depth of field, we find the deviation in D from the nominal position D_0 where it is in best focus. Since the local optics collect light across the entire aperture, the focus is degraded with deviation in D. By the lens law,

$$1/f = 1/A + 1/B, \text{ and } 1/g = 1/C + 1/D.$$

Using the magnification factors M and N, we solve for B and D to obtain

$$B = (M+1)f, \text{ and } D = (N+1)g, \text{ or } D = (1/g - 1/C)^{-1}.$$

Substituting E and B for C, we obtain

$$D = \left[\frac{1}{g} - \frac{1}{(E-B)}\right]^{-1} = \left[\frac{1}{g} - \frac{1}{D_0/N_0 + (M_0 - M)f}\right]^{-1}.$$

This establishes the desired relationship between A and D in terms of the magnification M. The above expression for D shows that, as the object moves to infinity, the total movement in the primary focal plane is $M_0 f$. The total movement in the secondary focal plane is further reduced from this value, which results in a wider range of focus over conventional imaging. In our example, the movement in the primary focal plane is 100μm for an object distance of 1m to infinity. This translates into a mere 1.5μm deviation in D. The magnification factor N varies from 1/4 to 1/16. It is clear that, even with wide local apertures, the system is adequate for measuring depth while maintaining focus. Note that although objects remain in focus, the effective spatial resolution is decreased due to demagnification.

B. Depth Extraction

To obtain an expression for depth, consider Fig. 4(b) and let L be the distance between a pair of apertures and Δ be the displacement of the image between apertures. We estimate the distance A from Δ. Again, E is adjusted to meet the desired magnification N_0 according to the other fixed parameters. The geometry of the configuration from the sensor to the primary focal plane gives:

$$C/L = D_0/\Delta.$$

Using the lens law for A as a function of B and making the substitution $B = E - C = B_0 + C_0 - C$, we obtain

$$A = \left(\frac{1}{f} - \frac{1}{B}\right)^{-1} = \left(\frac{1}{f} - \frac{1}{B_0 + C_0 - C}\right)^{-1}.$$

Solving for A in terms of Δ gives the depth equation

$$A = \left[\frac{1}{f} - \frac{1}{(M_0+1)f + D_0/N_0 - D_0 L/\Delta}\right]^{-1}.$$

A characteristic of this sensor is that the amount of depth information available is a strong function of the object distance (the closer the object, the higher the depth resolution). We can quantify this by solving for Δ in terms of M, which gives

$$\Delta = \frac{D_0 L}{(M_0 - M)f + D_0/N_0}.$$

As M increases, Δ rapidly approaches its limit of $D_0 L/(M_0 f + D_0/N_0)$.

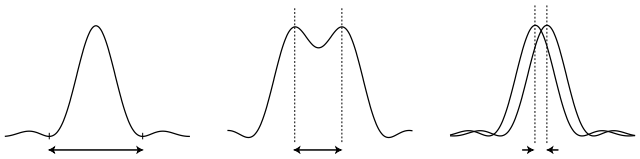
The rate of change in Δ with A, i.e., $\partial\Delta/\partial A$, can be computed as a function of $\partial B/\partial A$ and $\partial\Delta/\partial C$. Setting $\partial C = -\partial B$ at the focal plane, it can be shown that

$$\partial\Delta/\partial A \approx -\frac{f^2}{A^2}\frac{DL}{C^2} \longrightarrow \partial\Delta/\partial A \approx -M^2 N^2 \frac{L}{D}.$$

For example, if assuming 0.5μm pixel pitch, the displacement between apertures can be estimated to within 0.5μm resolution. Further, assuming $L/D = 2$, the incremental depth resolution ∂A is approximately 4cm at $A_0 = 1$m and 4mm at $A_0 = 10$cm. Decreasing pixel size allows for more accuracy in $\partial\Delta$ leading to higher depth resolution.

C. Spatial Resolution and Pixel Size

Clearly, the spatial resolution of our system is limited to the total number of pixels mnk^2. However, to establish overlapping fields of view, we set the magnification factor of the local optics to $N < 1$. Since each pixel is projected up to the focal plane by a factor of $1/N$, spatial resolution is reduced by $1/N^2$. Thus, the total available resolution is $\approx mnk^2N^2$.



(a) $W \approx \lambda/\text{NA}$, Diffraction limited (b) $W \approx \lambda/2\text{NA}$, Rayleigh Criterion (c) $W < \lambda/\text{NA}$, Multi-aperture

Fig. 5. Spot size comparison.

In our example, we assume a 16×16 array of 0.5μm pixels with a magnification factor of $N_0 = 1/4$. Thus, the maximum resolution is 16 times greater than the aperture count itself but 16 times lower than the total number of pixels.

The actual spatial resolution is limited by optical aberrations and ultimately by diffraction. The minimum spot size W for a diffraction limited system is $\approx \lambda/\text{NA}$ (see Fig. 5(a)), where NA$= n_i \sin\theta$ is the numerical aperture of the local optics, n_i is the index of refraction of the dielectric and θ is the angle between the chief and the marginal rays. Using the Rayleigh criterion, the minimum useful pixel pitch is commonly assumed to be half the spot size (see Fig. 5(b)). Assuming $n_i \approx 1.5$ in the dielectric stack, NA can be about 0.5, which gives a spot size of $\approx 1\mu$m. Thus, scaling the pixel beyond 0.5μm does not increase spatial resolution. Although no further increase in spatial resolution is feasible beyond the diffraction limit, depth resolution continues to improve as long as there are features with sufficiently low spatial frequencies. Indeed, the disparity between apertures can be measured at smaller dimensions than set by the diffraction limit as illustrated in Fig. 5(c).

IV. Implementation

As discussed, the proposed image sensor consists of an $m \times n$ array of apertures each having a $k \times k$ pixel array. Here we briefly discuss the design of the pixel array and how readout is performed.

The fact that our image sensor is composed of many small pixel arrays that can have gaps between them makes it feasible to scale pixel size beyond current sensor design limits. Specifically, we use a frame transfer charge-coupled device (FT-CCD) array with small pixels at each aperture. Because of the small array size, acceptable readout performance can be achieved with a modest charge transfer efficiency. By lowering the requirement on charge transfer, such a CCD becomes feasible to implement in CMOS with minor process modifications. As shown in Fig. 6, the pixel array is divided into two sections, a light sensitive CCD array of $k \times k$ pixels and a light shielded CCD array of $k \times k$ storage cells. The pixels in the entire image sensor are set to integrate simultaneously via global control. Such *global shuttering* is important here because of the need for highly accurate correspondence between apertures in extracting depth. After integration, the charge from each pixel array is shifted into its local frame buffer and then read out through a floating diffusion node via a follower amplifier. A correlated double sampling scheme is used for low temporal

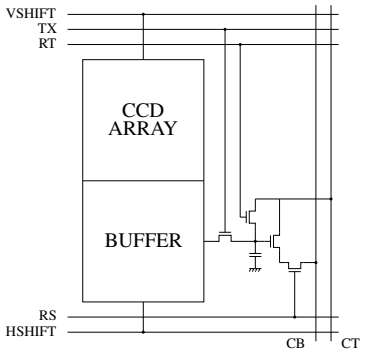


Fig. 6. Single aperture array with readout.

and fixed-pattern noise. Global readout is performed using hierarchal column lines similar to hierarchal bit/word lines used in low-power SRAM (such a scheme is not desirable in a conventional image sensor because it causes optical non-uniformity). Digitization is performed using column-level ADCs for fast readout or on-chip parallel processing.

V. Conclusion

The multi-aperture image sensor extracts a depth map of the scene by solving the correspondence problem between multiple views of the same points in the primary focal plane. The spatial resolution of the system is shown to be greater than the aperture count itself and governed by the magnification of the local optics and pixel size. The amount of depth resolution available is shown to increase with decreasing pixel size while the 2D spatial resolution remains limited.

In addition to providing depth information, the multi-aperture image sensor architecture can be used to improve the performance of color imaging. It can be shown that employing a per-aperture color filter array (CFA), instead of the conventional per-pixel CFA, can largely eliminate the color aliasing and crosstalk problems resulting from the large dielectric stack heights relative to pixel size in sub-micron CMOS image sensors.

Acknowledgment

Keith was supported by a Fannie and John Hertz Foundation Fellowship. Abbas was supported under DARPA Microsystems Technology Office Award No. N66001-02-1-8940.

References

[1] S. Marapane and M. Trivedi, "Region-based stereo analysis for robotic applications," *IEEE Trans. Syst., Man, Cybern.*, pp. 1447–1464, Nov./Dec. 1989.

[2] C. Niclass, A. Rochas, P.-A. Besse, and E. Charbon, "Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes," *IEEE J. Solid-State Circuits*, pp. 1847–1854, Sept. 2005.

[3] E. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[4] J. Tanida *et al.*, "Thin obervation module by bound optics (TOMBO): Concept and experimental verification," *Applied Optics*, vol. 40, pp. 1806–1813, Apr. 2001.