Real-Time Depth from Focus on a Programmable Focal Plane Processor

Julien N.P. Martel, *Student Member, IEEE,* Lorenz K. Müller, *Member, IEEE,* Stephen J. Carey, *Member, IEEE,* Jonathan Müller, Yulia Sandamirskaya, *Member, IEEE,* and Piotr Dudek, *Senior Member, IEEE,*

Abstract—Visual input can be used to recover the 3D structure of a scene by estimating distances (depth) to the observer. Depth estimation is performed in various applications, such as robotics, autonomous driving, or surveillance. We present a low-power, compact, passive and static imaging system that computes a semi-dense depth map in real time for a wide range of depths. This is achieved by using a focus-tunable liquid lens to sweep the optical power of the system at a high frequency, computing depth from focus on a mixed-signal programmable focal-plane processor. The use of local and highly parallel processing directly on the focal plane removes the sensor-processor bandwidth limitations typical in conventional imaging and processor technologies and allows real-time performance to be achieved.

Index Terms—depth from focus, tunable lens, vision-chip, focal plane processing.

I. INTRODUCTION

IN MANY practical applications such as autonomous navigation, action recognition, or environment mapping, knowing the three-dimensional structure of a scene is necessary. The 3D structure can be inferred from visual input and serves, for instance, to avoid obstacles, find the spatial extensions of objects and disambiguate their poses, or get a proper length scale in the environment.

There exist a number of cues that can be used to estimate distances from an observer in a scene. LIDARs [1] and SONAR [2] are active systems that measure the time-of-flight of an emitted and reflected light or sound wave. Structured-light [3], another active approach, illuminates the scene with a known pattern and infers shape and/or depth from the deformations of this pattern. Among passive approaches, we find stereo vision [4], in which matching points are identified, observed by two separated cameras with known optics and separation baseline; the direction of the two rays pointing towards a given point in the environment allows a simple geometric estimation of the depth. Optical flow inferred from monocular video can also be used for depth estimation [5].

These different systems and the underlying algorithms have different strengths and limitations and thus vary in their domain of applicability. For instance, structured light works best in low-light settings indoors. The reason for this is that the projected patterns are most visible indoors, where there is

J.N.P. Martel, L.K. Müller, J. Müller and Y. Sandamirskaya are with the Institute of Neuroinformatics, University of Zurich and ETH Zurich, CH-8037 Zurich, Switzerland, (e-mail: jmartel@ini.ethz.ch).

S.J. Carey and P. Dudek are with the School of Electrical & Electronic Engineering, The University of Manchester, M13-9PL Manchester, United-Kingdom
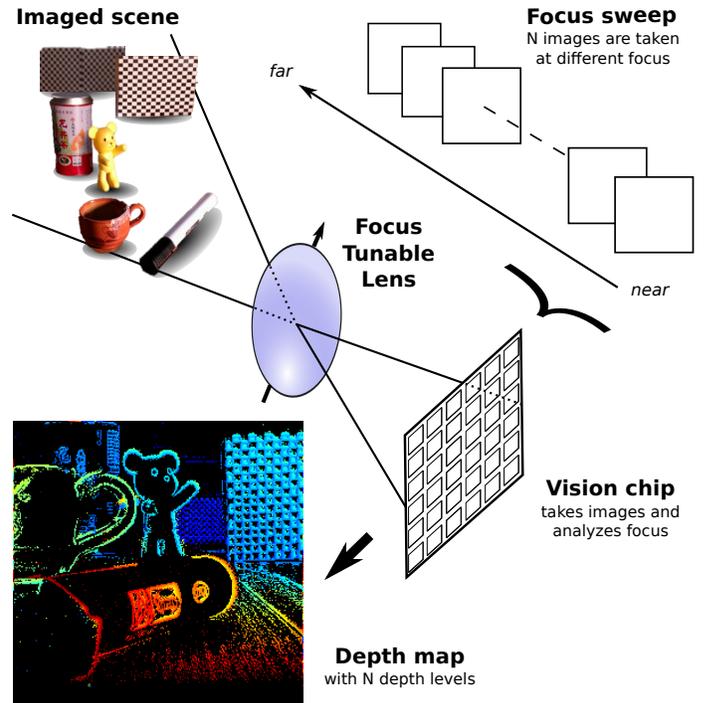
Fig. 1. Schematic of the depth from focus system. A vision chip images the scene through a focus tunable lens, with rapidly changing optical power. Within a 'focus sweep', the vision chip analyzes the sharpness of each pixel in N images. This yields a depth map with N depth levels.

no interference with the outdoor sunlight or sources of heat. In the case of stereo, the baseline of the system, i.e., the distance between the two cameras, determines the maximal depth range for a given resolution (and a bigger baseline reduces the compactness of the overall system); optical flow based approaches require only a standard video camera, but are less suitable for resource-limited platforms because they are generally computationally costly.

In this work we present a monocular depth imaging system. The system is passive, i.e. it does not actively emit energy into the scene; it is static, i.e. it does not require motion of the camera; and it operates on a low power budget, due to the highly parallel nature and locality of its underlying computational architecture and use of efficient analog computing elements. All these properties contribute to making the system compact and energy efficient, which is desirable for embedded applications.

The system's key operational principle is the evaluation of 'sharpness' of a given image region. While changing the optical power of an optical system with a large aperture,
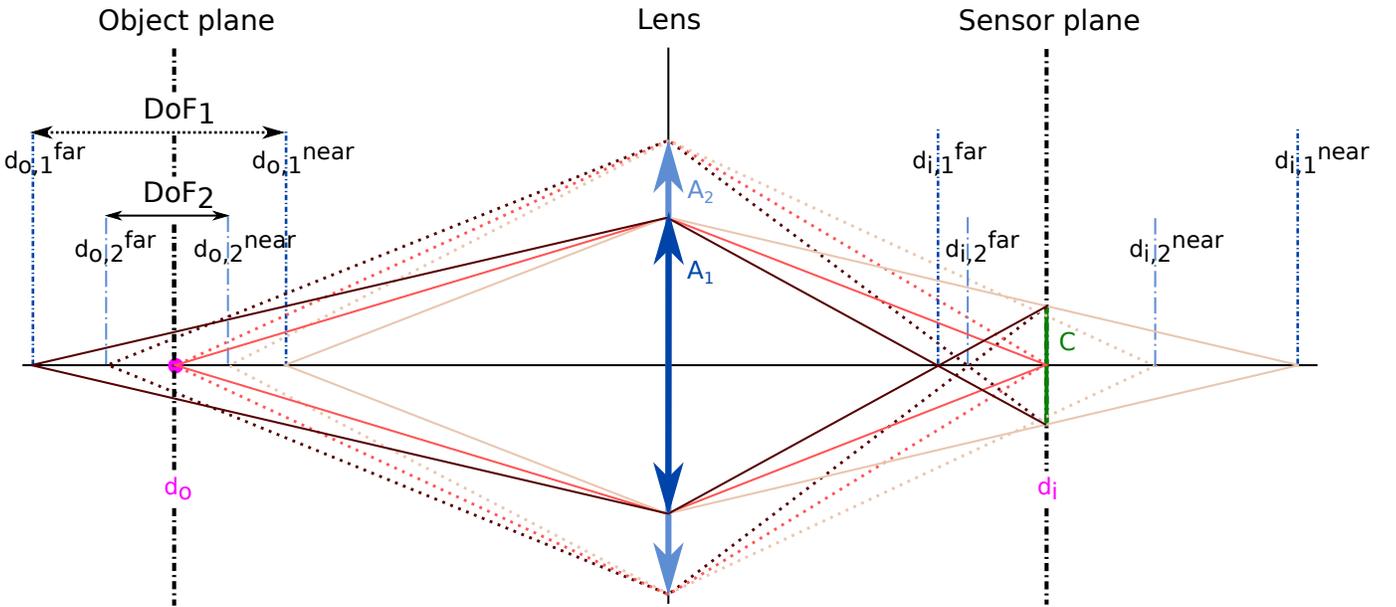
Fig. 2. An illustration of the depths of field $DoF_{1/2}$ and the near/far limits of focus $d_o^{\text{near/far}}$ for two apertures $A_{1/2}$ using a convex-lens in front of a point. The size of a pixel on the image plane/sensor plane defines the circle of confusion C that sets the upper limit of resolution (Figure adapted from [6]).

objects at different depths fall in and out of focus, due to the limited depth of field of such systems. By identifying the optical power at which an image region is maximally in focus, or sharpest, we can infer the depth of the region (see Fig. 1 for an illustration). Since sharpness can only be estimated in textured regions (containing high spatial frequencies), we only recover depth in such regions and thus obtain sparse (a.k.a semi-dense) depth maps.

We realized the algorithm for evaluation of 'sharpness' on a mixed-signal, low-power vision chip (focal plane processor array), SCAMP-5 [7], which we instrumented with a focus-tunable lens [8]. We can configure the lens to sweep the focus back and forth at high frequency (up to a kHz). Multiple image frames are acquired during the focus sweep by the vision chip. For each pixel in each frame, the SCAMP processor analyzes sharpness (in the form of a local contrast measurement). For any given pixel, the frame with the maximal sharpness is tagged with the focal power that corresponds to the optical parameters that were set when the frame was captured. This focal power can then be associated with a depth value as we elaborate in Sections II and III.

Using a conventional camera and a central processor would incur a prohibitive communication overhead, as multiple frames need to be processed for a single depth map. We can circumvent this limitation by making use of the ultra-high bandwidth between the pixel-parallel focal-plane processor and its sensor data. The photosensor of each pixel is directly coupled to the processing element. The sharpness analysis of each image is carried out on-chip in the focal plane and in parallel for all pixels, and only the final computed depth map is transmitted off-chip. Further details about the system's implementation are given in Section IV. We show in Section V that this allows us to obtain sparse depth images with 32 depth levels at > 25FPS and that we can trade-off frame-rate for depth resolution, up to a fundamental limit posed by the optics

and the finite aperture of the system.

II. PRINCIPLES AND LIMITATIONS OF DEPTH FROM FOCUS

Depth from focus is based on the idea that using lenses with large apertures, a point in the scene at a certain distance from the observer only appears in focus for a certain optical-power (the inverse of the focal length). In other words, for a system that has a shallow depth-of-field, a small range of depth in object space (ultimately a single point) is in focus in the image for a given optical power.

A. A Simple model of focus with geometric optics

The idea of estimating depth from focus lies in geometric optics and can be dated back at least as far as the work of Horn [9]. Consider a thin convex lens with an infinite circular aperture, and rays forming small angles with the optic axis (OA). For a focal length f – the distance at which all incoming parallel rays converge – a point on a plane perpendicular to the OA in front of the lens at distance d_o, forms an image behind the lens on a plane perpendicular to the OA placed at distance d_i according to the 'Gaussian lens' equation:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}. \tag{1}$$

It is convenient to define the optical power δ as the inverse of the focal length: $\delta = f^{-1}$ and write, with inverse distances $\hat{d}_o$ and $\hat{d}_i$ that:

$$\hat{d}_o + \hat{d}_i = \delta. \tag{2}$$

If one places an imaging sensor orthogonal to the OA at distance d_i, the system is said to be "in-focus for the object points at distance d_o". If an object point is translated to $d'_o \neq d_o$, keeping d_i fixed, then the light originating from the object point is redistributed according to the lens's point spread function and the image point is blurred.
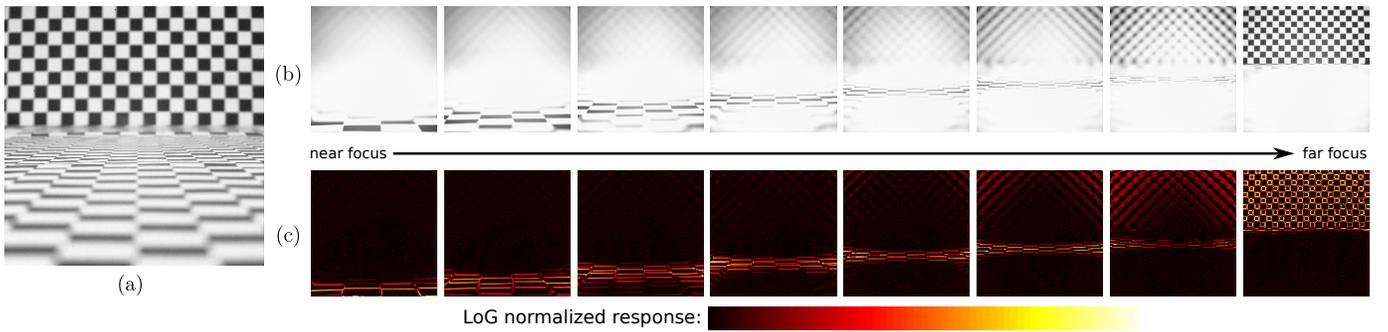
Fig. 3. Demonstration of the depth-from-focus principle: (a) shows the setup for the experiment – a checkerboard pattern is imaged (here the picture is taken with a very narrow aperture so that the whole scene is in focus); (b) shows sample images taken during a focus sweep at different optical powers from near to far focus. These are taken with a wide aperture to minimize the DoF; (c) shows the Laplacian of Gaussian (LoG) response for each sample image.

Thereby, if an object forms a sharp image on an image plane placed at d_i, it can be related, given the optical power δ of the lens, to the unique distance in object space d_o at which it is in focus. Thus, its depth (i.e. its distance to the lens) is:

$$\hat{d}_o = \delta - \hat{d}_i. \tag{3}$$

This gives us a way to compute the distance of a point from the observer given the configuration of the system: the distance between the lens and the sensor, d_i, and the optical power, δ. Hence, all the points that are seen in focus in the image can be inferred to be at depth d_o.

For a depth imaging system, we are not solely interested in objects lying at a single distance from the lens, but in real scenes containing objects that may be placed at different distances. Keeping in mind the simple model described previously, to bring in focus different object planes and get access to a range of depth values, one has to vary at least one of the three parameters d_o, d_i, or δ.

- *Varying d_o, the relative lens-object distance* can be achieved by "translating" the scene towards/away-from the camera. As it translates, objects at different depths fall in focus. Taking such an action is impractical or impossible in many situations, though it was the first kind of systems to be developed [10]. On the other hand, this has the great advantage of keeping magnification constant. Magnification is the scale (the zoom) at which the imaging system produces images. It is defined as $m = \frac{d_i}{d_o} = \frac{f \cdot d_o/(d_o-f)}{d_o} = \frac{f}{d_o-f}$ (using the thin lens equation Eq. (1) to express d_i as a function of d_o), from which it is clear that if f is constant, m is also constant. A constant magnification factor through the whole sweep is important, as otherwise artifacts can be observed (such as in Figure 6d, in which the changing magnification confers an ability to see "through" the case of the camera in the image).
- *Varying d_i, the relative lens-sensor distance* can be achieved by modifying the relative distance between the lens and the imaging plane. This is the principle of previous works such as in [11] and [12]. An issue lies with the very small scale of such translations (on the order of micrometers) and the difficulty to perform them accurately at high-speed; it is thus impractical to implement. Note that in this case, as d_i is varied, even though f is constant, the magnification is not

kept constant, since the magnification can also be expressed as $m = \frac{d_i}{f \cdot d_i/(d_i-f)} = \frac{d_i-f}{f}$.
- *Varying δ, the optical power* of the lens (or multi-lens equivalent) can be achieved using a mechanical system. Camera objectives equipped with motorised focus rings are widely available nowadays. Computer vision has traditionally addressed the problem of recovering Depth from Focus/Defocus offline, using focal stacks (images taken at different focus) for static scenes [13]. However, if one aims at designing a high-speed depth imager, the slowness of such mechanical systems becomes an inconvenience. In this case, even if d_i remains fixed, magnification is also not kept constant as f is being varied. This can be addressed by using an optic system that is telecentric.

The system we present in this work falls into this last category. Its novelty is to combine a focal plane processor array with a focus-tunable liquid lens. Although such lenses have been used to perform fast autofocus (e.g. to perform focus positioning when coupled with a depth finder) we suggest using them in a continuously sweeping mode, allowing focus change at frequencies of up to a kilohertz.

B. Depth of field limitations

To better understand the design challenges and limitations of the system presented here, a more detailed model of focus needs to be introduced. Specifically, the aperture of a real optical system is not infinite, but has a maximal – and assumed to be fixed – diameter A. Furthermore, an imaging system does not present an infinite resolution, but is ultimately limited by the size of its pixels, below which it is not possible to distinguish focus. This defines the circle of confusion C. Taking these properties of the real optical system into account, one can observe that instead of a single point, a range of points in front of the lens project within the circle of confusion and therefore all fall in focus, as illustrated in Figure 2. The limits of the range in depth that falls into focus when focusing at d_o – between the near-focus d_o^{near} and far-focus d_o^{far} – can be derived.

Looking at Figure 2 and using the intercept theorem, we write the relation between the projection behind the lens of

the near limit of focus d_i^{near}, the sensor's plane position d_i, the circle of confusion C, and the aperture of the system A:

$$\frac{C}{A} = \frac{d_i^{\text{near}} - d_i}{d_i^{\text{near}}} \Longleftrightarrow d_i^{\text{near}} = \frac{d_i}{1 - C/A}, \quad (4)$$

similarly for the projection behind the lens of the far limit of focus d_i^{far}:

$$\frac{C}{A} = \frac{d_i - d_i^{\text{far}}}{d_i^{\text{far}}} \Longleftrightarrow d_i^{\text{far}} = \frac{d_i}{1 + C/A}. \quad (5)$$

Using the Gaussian Lens Equation (1), and applying it to the pair d_o^{near} (in object space) and d_i^{near} (in image space), we have:

$$\frac{1}{d_o^{\text{near}}} = \frac{1}{f} - \frac{1}{d_i^{\text{near}}}, \quad (6)$$

and for the pair $\{d_o^{\text{far}}, d_i^{\text{far}}\}$:

$$\frac{1}{d_o^{\text{far}}} = \frac{1}{f} - \frac{1}{d_i^{\text{far}}}. \quad (7)$$

Inserting the values of d_i^{near} and d_i^{far} using Equations (4) and (5), we obtain the limits in object space of near and far focus:

$$\frac{1}{d_o^{\text{near}}} = \frac{1}{f} - \frac{1 - \frac{C}{A}}{d_i} = \frac{d_i - f(1 - \frac{C}{A})}{f \cdot d_i} = \frac{A(d_i - f) + fC}{Afd_i} \quad (8)$$

and

$$\frac{1}{d_o^{\text{far}}} = \frac{A(d_i - f) - fC}{Afd_i}. \quad (9)$$

This gives us the general form of the near and far limits of focus in object space given the parameters of our system $\{A, d_i, f, C\}$ as:

$$d_o^{\text{near/far}} = \frac{d_i Af}{\pm Cf + A(d_i - f)}. \quad (10)$$

Hence, we can calculate a range of points in focus between near and far limits: this is called the depth of field DoF and is given by:

$$DoF = d_o^{\text{far}} - d_o^{\text{near}} = \frac{-2d_i Af^2 C}{C^2 f^2 - (d_i - f)^2 A^2}. \quad (11)$$

We have now presented equations that describe how the near/far limits of focus and the depth of field can be expressed given all the intrinsic parameters of our system. We use these formulas to recover the range of depths corresponding to points in focus in the image given a particular optical power/focal setting f of the focus-tunable lens. Note that again using the Gaussian lens equation (1), one can re-derive Equations (10) and (11) as a function of d_o. These describe what the near/far limits of focus, and depth of field are when focusing at a certain point d_o. Thus an alternative form of Equation 10 as a function of d_o is:

$$d_o^{\text{near/far}} = \frac{d_o Af}{Af \pm C(d_o - f)}. \quad (12)$$

And the following expression calculates the depth of field when focusing at d_o:

$$DoF = \frac{2d_o AfC(d_o - f)}{A^2 f^2 - (d_o - f)^2 C^2}. \quad (13)$$

The interpretation of these equations yields important results concerning the limitation of our system. In particular, how resolution is limited due to large depth of fields at long distances: (a) as d_o increases the DoF becomes very large, since C is orders of magnitude smaller than $A f$, which means that the system cannot resolve depth past a certain distance – the 'hyperfocal distance'; (b) the DoF depends linearly on the inverse of the aperture, a larger aperture is better. Note that an infinite aperture as described in Section II-A would yield an infinitely small DoF and infinitely high resolution.

It is useful to note that having a large aperture is conceptually not very different from a stereo-vision system. In stereo-vision, two cameras at physically different positions and different angles are looking at a scene. The cameras share some of their field of view and thus the images they produce overlap. For a given point in the scene, the two cameras project it on their image plane at different positions, and there is a displacement between the point as seen in the two images they produce. This disparity between the two projections of the point on the images allows the point to be triangulated in three-dimensions and thus inference of distance (depth) with respect to the observer. The aperture is to depth-from-focus what the baseline – the distance between the two cameras – is to stereo-vision. With a very large baseline between two cameras in a stereo vision setup it is easy to triangulate the point in three dimensions with high precision. Similarly, the larger the aperture, the better the capability to resolve depth, since the depth of field is shallower.

So why use depth from focus? An advantage of depth from focus is that the optic centers of the two "virtual" cameras are aligned and no complex realignment of two images and search for matching points has to be performed.

III. AN ALGORITHM TO RECOVER DEPTH FROM FOCUS

The basic principle of our algorithm is to change the imaging system over time and to make note of the time at which a given object appears maximally sharp. We then associate this time to the corresponding system state and infer the distance to the object from the geometric optics as outlined in the previous section.

If we sweep the optical power, δ, of the lens periodically through time, $\delta : t \to \delta(t)$, objects lying between the closest near limit of focus ($d_{o,\min}^{\text{near}}$ associated to the highest reachable optical power $\delta_{\max}$) and the furthest limit of focus ($d_{o,\max}^{\text{far}}$ corresponding to the lowest reachable optical power $\delta_{\min}$) will be brought in and out of focus during this focal sweep. Each object will appear sharp in the image for some time interval given by the speed of the sweep and the DoF at the object distance. We sample this focal sweep by taking N individual images at equal time intervals. Within this sweep we determine for each pixel in which image it was maximally sharp.

In previous work [14] we demonstrated that the Laplacian of Gaussian (LoG) can be computed efficiently on the cellular processor array architecture we use. SCAMP-5 [7] has a dedicated diffusion network that enables Gaussian blur in a single clock cycle. The fast neighbour to neighbour communication allows for convolutions with a discretized Laplacian filter in

---

Algorithm 1 The algorithm implemented on-chip

Require: R_I, R_L, $R_L^{\max}$, R_M: four register arrays (image, focus response, max focus response and "frame" index)
Require: R_{pix}: the image acquisition array (the photo-sensors)
Require: N, ΔT: two parameters (number of depth levels and exposure control)
1: $R_M \leftarrow N$ ▷ Initialize default index to 'far' (background)
2: $R_L^{\max} \leftarrow 0$ ▷ Initialize the max focus metric value to 0
3: $n \leftarrow 1$
4: **while** $n < N$ levels **do**
5:     $R_I \leftarrow R_{\text{pix}}$ ▷ Load current image
6:     $R_{\text{pix}} \leftarrow 0$ ▷ Start light integration for next image
7:     $R_L \leftarrow f(R_I)$ ▷ Compute focus metric
8:     **if** $R_L > R_L^{\max}$ **then** ▷ Compare focus to max focus
9:         $R_L^{\max} \leftarrow R_L$ ▷ Store max focus
10:         $R_M \leftarrow n$ ▷ Store iter. num.
11:     **end if**
12:     wait ΔT ▷ Control R_{pix} exposure
13:     $n \leftarrow n + 1$
14: **end while**
15: **if** $R_L^{\max} < \theta$ **then**
16:     $R_M \leftarrow 0$ ▷ Discard pixels w/. low maxima
17: **end if**
18: trigger depth-frame readout of R_M

---

a dozen of cycles. The LoG can be computed in a pipelined manner, its evaluation takes place while acquiring the following image. In fact, the computation of the LoG is so fast using the diffusion network and neighbour communication, that we need to introduce a waiting time after its computation to finish the exposure of the subsequent frame, as illustrated in Figure 5 and explained in a greater detail in Section IV-C.

Within a focus sweep, we keep track of the maximal sharpness any given pixel achieved, as well as the index of the respective image in the sweep; thus we obtain the index of maximal sharpness in a single pass. Additionally, we compare this maximal sharpness to a user-defined threshold θ. Wherever the scene contrast is too weak for the focus measure to exceed the threshold, the data is marked as invalid. From the N exposures taken during a focus sweep this algorithm yields a single, sparse depth frame with N depth levels.

The user-defined threshold θ discards noise and ensures that spurious maxima (for instance, arising from image magnification) are suppressed. In practice, a small image magnification arises when using non-telecentric optics, because when focusing, the whole focal-length of the system is slightly changed, and consequently FoV and magnification are also slightly changed. This effect is known as "breathing" and thresholding is a simple way to account for it computationally[1]. The user-defined threshold also suppresses noise coming from strong contrast of, e.g., specular surfaces.

The threshold is determined empirically, but it could be estimated, for example, by a statistical method (i.e. be learned) to match a ground-truth depth-map if such is available. In

---

[1]It could be accounted for optically by using a telecentric back-lens

our experiments, we have set the threshold to 15% of the maximal possible sharpness (a black pixel surrounded by white neighbours) and this has not been changed across scenes.

The algorithm is summarized in Alg. 1. More formally the algorithm can be described as follows.

During the focus sweep, we acquire a set of images I_n, $n \in \{1, 2, ...N\}$ with N the number of desired depth levels. We index the pixel at coordinates (x, y) with $p(x, y)$.

For each image, we compute a focus metric L_n using an operator f. We choose our focus metric to be the response of the image to the filtering with a Laplacian of Gaussian (LoG, see Fig. 3 for example responses), which is a "high-pass" filter:

$$L_n = f(I_n) = I_n * \text{LoG}. \tag{14}$$

The Laplacian of Gaussian (LoG) is defined by:

$$\text{LoG}(x, y) = \Delta G_\sigma(x, y) = \frac{x^2 + y^2 + \sigma^2}{\sigma^4} e^{-(x^2+y^2)/2\sigma^2}, \tag{15}$$

where Δ is the Laplacian operator and G_σ, a Gaussian with standard deviation σ. In practice we compute an approximate LoG by first approximating a Gaussian blur on $I_n(p)$ and then using a discretization of the Laplacian on the resulting blurred image using the classical 4-points stencil:

$$LoG = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix} * G_\sigma(x, y). \tag{16}$$

Then, to obtain a depth map, we first compute an index map M by finding for each pixel the index that maximizes the response of the LoG (our criterion for sharpness):

$$\text{M}(p) = \underset{n \in \{1...N\}}{\arg \max} L_n(p). \tag{17}$$

This depth map is the output of our algorithm. Finally, we can associate to each pixel a depth value in a depth map D given the sampling scheme used during the sweep $S : i \to \delta$ that specifies how the index of a sampled image relates to the optical power it was taken with during the focus sweep, and the relation between optical-power and depth $H : \delta \to d$:

$$D(p) = H \circ S \circ \text{M}(p). \tag{18}$$

In our particular case we use a uniform sampling scheme, that is for N images (corresponding to N depth levels) we distribute our samples equally during the focus sweep ranging from $\delta_{\min}$ and $\delta_{\max}$. For an index i the sampling function relating index to optical power is then:

$$S(i) = \delta_{\min} + i \cdot \frac{\delta_{\max} - \delta_{\min}}{N}. \tag{19}$$

In the simplest case the relation H between optical power and depth is given by Equation (1).

Notably, Equation (11) suggests that equally spaced sample images are non-optimal. To achieve maximal frame-rate and number of depth levels, one should space the sampled images such that their corresponding DoFs are non-overlapping. To achieve this one would either need a non-linear sampling procedure (cf. Section II-B) for a linear focus sweep, or a focus that depends on time in a non-linear manner. The former
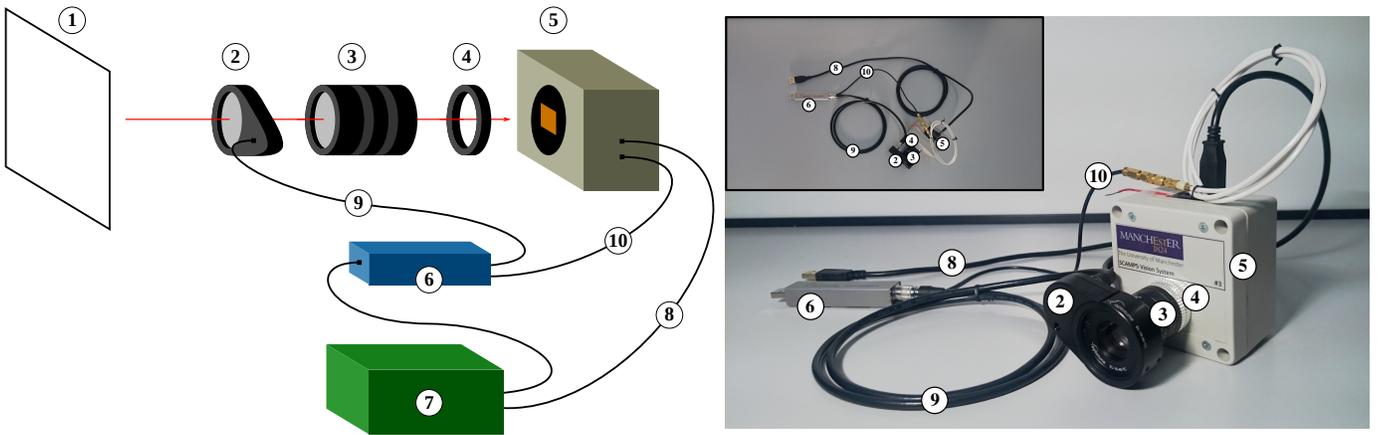
Fig. 4. A diagram with the top and front views of the assembled system: (1) object plane, (2) focus tunable lens, (3) fixed objective lens, (4) distance ring, (5) SCAMP-5 sensor and case, (6) lens current driver, (7) embedded computer for frame readout, (8) USB-Cable between SCAMP-5 and computer, (9) lens cable, (10) frame trigger cable.

has the drawback of limiting the exposure time available for near-focus samples and a non-linear focus sweep cannot be implemented as easily on the focus-tunable lens. Because of these difficulties, we do not address this issue further in the scope of this work.

IV. DESCRIPTION OF THE SYSTEM AND IMPLEMENTATION

A schematic diagram of the full system and its components is given in Fig. 4. We use a pixel-parallel processor array vision chip [7] placed behind a fast focus-tunable lens [8] to perform the algorithm presented in Section II.

A. Optical system

Our optical system comprises three different components: (a) *An objective* composed of a multi-lens system whose focus can be manually tuned is attached to the vision chip, we refer to it as the back-lens. (b) *The focus tunable lens* is attached to the front of this objective, separated by (c) *a lens of negative focal length*, referred to as an offset lens, whose purpose is to bring the focus tunable lens into the focal range of the back-lens. Alternatively, a concave tunable lens could be used to replace (b) and (c). The tunable lens we use is a liquid-lens (Optotune). Its working principle relies on a liquid placed inside a polymer membrane that is free to move in and out a peripheral reservoir, and thus change the inner membrane's curvature, hence the lens system's optical power [8].
Keeping the configuration of the front tunable lens and its offset lens fixed, the choice of the back-lens determines the working range of the system as well as its field of view. We performed experiments with a 13.5mm and 25mm lens, yielding a field of view of $41°$ and $23°$ and a working range of $[0.01, 3.5]$m and $[0.1, 11]$m respectively (we define the upper limit of this working range as the hyper-focal distance).

B. Focal Plane Processor Vision Chip

The focal plane processing device we consider here is the SCAMP-5 mixed-signal cellular processor array vision chip [7]. The device comprises an array of 256×256 processing

elements (PEs). A single PE includes a light-sensitive register, 6 local registers and a common register to share information with its 4-adjacent neighbours, and a comparator feeding an activity-flag latch. Analogue switched current techniques are employed to implement the PE: arithmetic operations are then performed without the need of a complex Artihmetic Logic Unit (ALU). The device is programmable: instructions are dispatched by an external global controller to all the PE cells. Each of them performs the given instruction on their local data thus implementing Single Instruction Multiple Data (SIMD) processing. The state of the activity-flag can enable or disable the operation of the cell preventing it from performing the instruction and so enables branching.

The benefit of using a vision chip device in a system with a high-speed tunable-lens is to prevent bandwidth issues related to the readout of all N images used in the reconstruction of a single depth-frame. Thanks to the very high bandwidth that exists between the sensor and processing part of such a device (the entire image is transferred to the processors in one clock cycle), the N samples can be easily processed *where* and *when* they are collected. The depth frame of N levels is then the only output for a time transfer cost of $\log_2(N) \cdot T_{\text{dig\_out}}$ (where $T_{\text{dig\_out}}$ is the readout time of a binary image array). This is very beneficial since the total transfer of each image to be processed externally would cost a time of $N \cdot T_{\text{ana\_out}}$ (where $T_{\text{ana\_out}}$ is the readout time of an analogue, or 8-bit, image array) and $T_{\text{ana\_out}} \gg T_{\text{dig\_out}}$.

C. Implementation and configuration of the system

The focus-tunable lens is driven by a current generator mapping linearly a range of $[0, 191]$mA to $[\delta_{\min}, \delta_{\max}] = [5, 10]$dpt of optical power translating to $[-1.5, 3.5]$dpt with the -150mm offset lens. We operate it with a triangular waveform and thus change linearly and periodically the optical power of the lens. As illustrated in Figure 5, a trigger from the current generator is synchronized with the sweep. The vision chip is slaved to this trigger and waits once the digital readout is performed until the next trigger arrives to start the imaging and processing of a new depth-frame. Since the up sweep
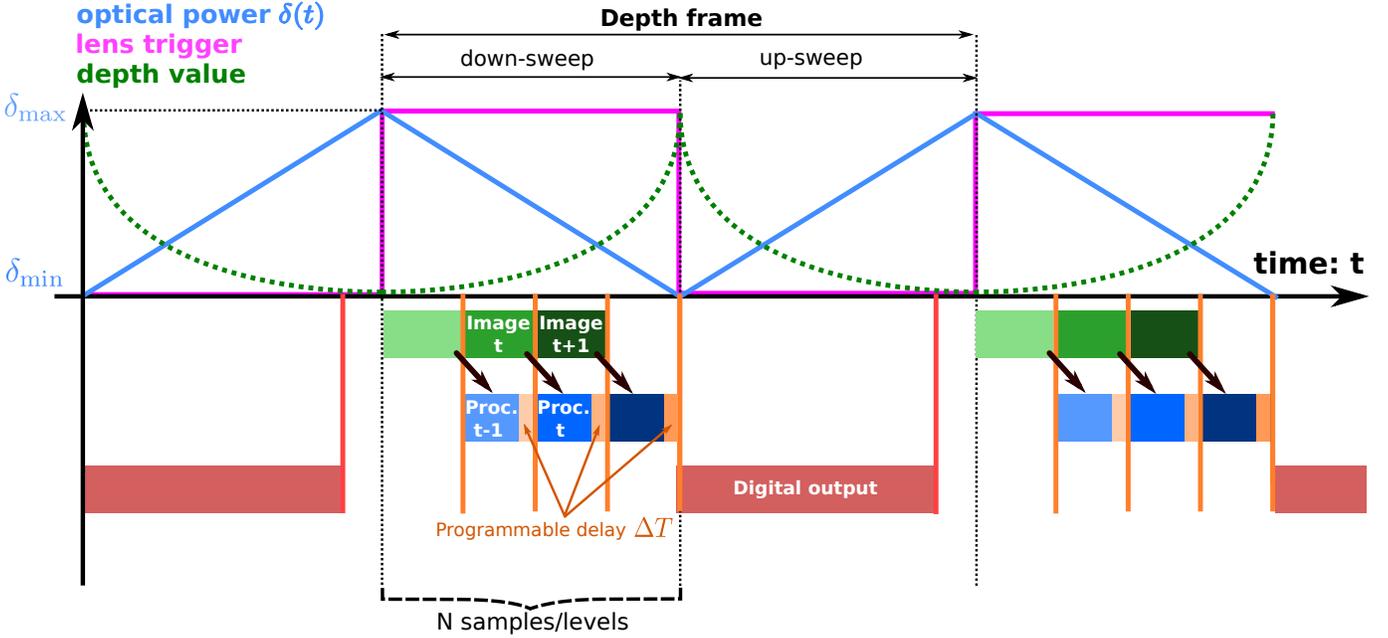
Fig. 5. A diagram illustrating timing in our depth from focus system and its pipelined imaging/processing operations.

(from near to far) contains redundant information with respect to the down sweep (from far to near), we solely use the down sweep to sample the focus whereas the time during the up sweep is used to read the digital depth frame out of the chip.

Our system is controlled by three parameters: The frequency of the lens ν, the number of depth levels (samples/images taken) N, and a programmable delay ΔT. We use this delay to pipeline imaging and sensing such that processing of an image $n \in \{1, \ldots N\}$ occurs when imaging the $n + 1$ sample. Since processing is very fast, we set the delay ΔT to increase the exposure time to a useful range. For the three parameters, only two degrees of freedom exist, since in order to maximally span the focal sweep (to get the maximal range of depth available) the following must hold:

$$\left(N \cdot (T_{\text{proc}} + \Delta T) = \frac{1}{2\nu}\right) \wedge \left(\lceil \log_2(N) \rceil \cdot T_{\text{dig\_out}} \leq \frac{1}{2\nu}\right), \ (20)$$

in which T_{proc} is a constant that measures the time needed on-chip to execute lines 2 to 9 of the algorithm we present in Algorithm 1 and $T_{\text{dig\_out}}$ a constant measuring the time to discard the low-confidence values and output a 1-bit plane digital register from the focal plane processor array (lines 13 to 16 of Algorithm 1). When the parameters ν, N, and ΔT satisfy parts of Equation (20), the lens frequency ν also determines the frame rate of the system. A trade-off between these three parameters has to be made: the more levels and the higher the frequency, the shorter the exposure's control ΔT must be.

V. RESULTS

In Figures 1, 6 and 7 we show images as output from our depth from focus system: up to N=64 levels are mapped using a heat-map look-up table. Colors correspond to the index of the level at which the maximal focus was achieved, as read out from R_M.

Figure 6 shows a number of exemplar scenes, analyzed by our depth-from-focus system, along with the respective depth map, obtained for objects at different distances from the sensor, in different illumination conditions, and at different parameters of the optical system and the algorithm.

Using a $f = 25$mm lens with a clear aperture of $A = 16$mm at distances $d_o = \{0.1, 0.5, 1.0, 5.0\}$m the corresponding depth of fields are $DoF = \{0.1, 4.0, 16.5, 50.0\}$cm. Note that depth resolution is physically limited: large DoF near the hyperfocal distance does not allow us to resolve depth accurately far away from the lens. In addition, having more levels for a fixed lens frequency ν degrades the focus metric since less light is captured. A higher N would thus only contribute in increasing the resolution near the lens and would consequently largely benefit from the use of a non-linear sampling scheme as suggested in Section III.

Our system requires under 1.9W of power, two thirds of which are drawn by the lens, the remaining third (633mW) by the focal plane processor. A third of these 633mW is the actual power drawn by the vision-chip, the rest being drawn by the instruction processing unit dispatching instructions to the vision-chip. The general characteristics of our system are summarized in Table I.

In addition to depth images, we can also record extended depth-of-field frames, as shown in Figure 7b. Extended depth-of-field frames are all-in-focus images: this is achieved by storing the current pixel value in the sample image for which the maximal focus is reached.

A. Trading-off the number of frames per second against the number of levels

We have demonstrated operation at up to 150FPS with N=16 levels under artificial lighting. Note, that in this setup, the vision chip captures images at $150 \times 16 = 2400$FPS which
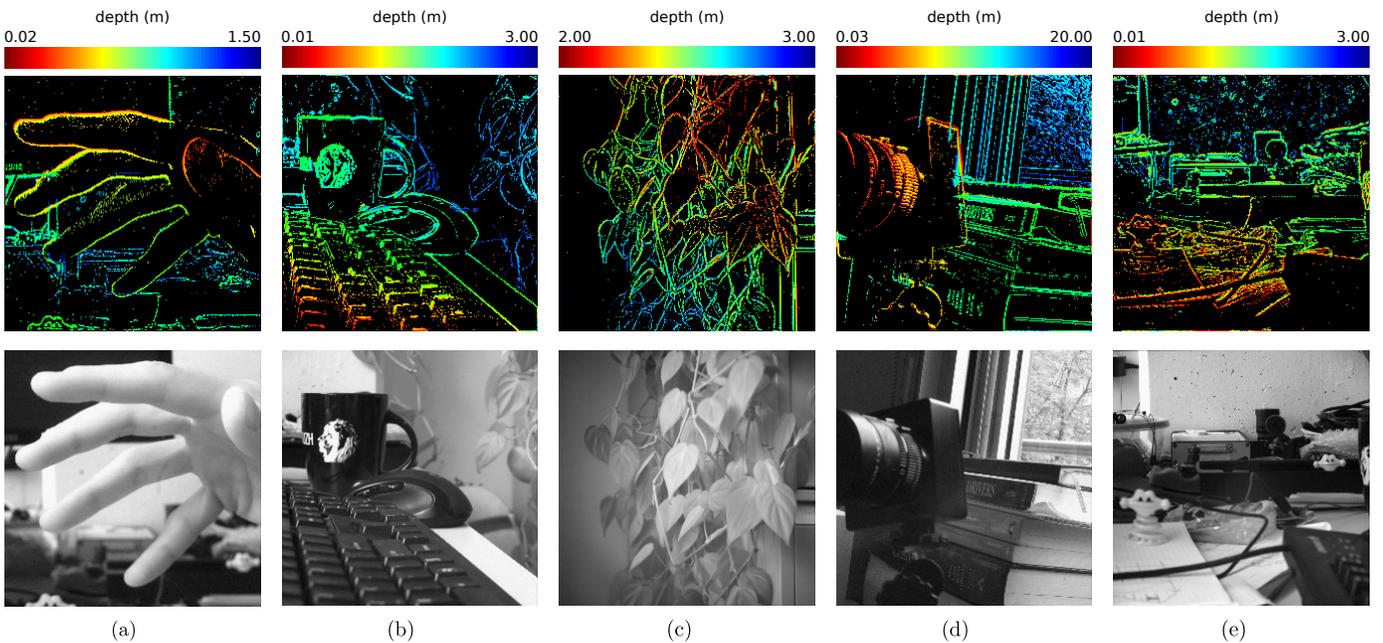
Fig. 6. Examples of scenes with different settings of the system (tuning the objective lens/back-lens to shift the depth at which the system operates and δ_{min} and δ_{max} to optimize the resolution within a certain range of depth. Note that the linear color scale is hyperbolic in distance. (a) the objective lens is set to focus at 20cm (before the thumb) and the tunable lens sweeps up to 1.5m. In (b) and (e) the objective lens is set to focus at 50cm and the tunable lens is also used in a concave mode (with negative focal lengths) to sweep from 10cm up to about 3m. In (c) we try to sweep a short depth range but further away: the objective lens is set to focus at about 2.5m and sweeps between 2m and 3m so that the leaves of the plants can be seen at different depths. Finally in (d), the objective lens is set to focus at about 4m and the focus tunable uses its whole range to sweep between 30cm to about infinity (as can be seen looking at the tree through the window).

corresponds to an exposure of 0.4ms. Under good lighting condition (10klx) the system runs with 32 levels at 25FPS and can be tuned, according to Eq. (20) for more levels and fewer frames per second, e.g. 64 levels at 16FPS. With the current system, it is possible to capture up to 128 levels. This number is limited by the number of digital registers available on chip. In low light conditions, our system is currently limited by the low fill factor of our pixel, and thus will need to be run at a lower frame rate, typically between $3-12$FPS indoor without artificial lighting.

Further speed improvements would be possible with faster readout circuitry, and improved light sensitivity. The latency of the system is dominated by the output of digital frames, which is about $T_{dig\_out} = 2$ms in contrast to a processing time of $T_{proc} = 56\mu$s that was measured for SCAMP-5 running at 10MHz. Increasing the number of levels N contributes to slowing down the system, mostly because of the increased time spent outputting data. To speed up the digital output, a sparse Address Event Representation (AER) [15] output might be an option; the sparsity of the frames (where sufficient contrast provides a reliable focus measure) is typically about $80-85\%$, and could be further increased using a higher threshold θ.

B. Trading-off the field of view against the resolution and the depth range

First, note that the following two quantities are traded-off: the field of view (that is changed by the focal length f of the whole system) and the depth range we can sweep with good resolution (depending on the hyper-focal distance). For a large

f, the field of view is narrower and the hyperfocal distance is larger. Thus, we propose two systems:

- a system using an objective lens with large field of view for high resolution in short-range imaging, with application on a table-top scenario, for instance;
- a configuration with a narrow field of view with resolution in a longer-range, with application on a mobile autonomous platform, for instance.

Details of the two configurations are given in Table I.

Furthermore, by changing where the objective lens/back-lens focuses we can shift the depth at which our system operates. In addition, by changing the range swept by the focus tunable lens, δ_{min} and δ_{max}, we change the range of depth swept in object space. In Figure 6, we illustrate how changing the configuration of the objective-lens/back-lens and the tunable lens affects these and allows us to obtain either a system that sweeps a large range of depth with low-resolution, a short range of depth at mid-distance with mid-resolution, or a short range of depth at short distance with high-resolution. We emphasize here once more that resolution for fixed aperture, fixed sensor's intrinsic parameters (pixel size and placement of the sensor behind the optics) depends on the focal length of the whole system and where focus is made (both allowing to derive the Depth of Field) as discussed in Section II-B.

C. Limitations of the system and outlook on how to address them

We showed that our system can generate sparse depth maps. In practical applications one might require dense maps,
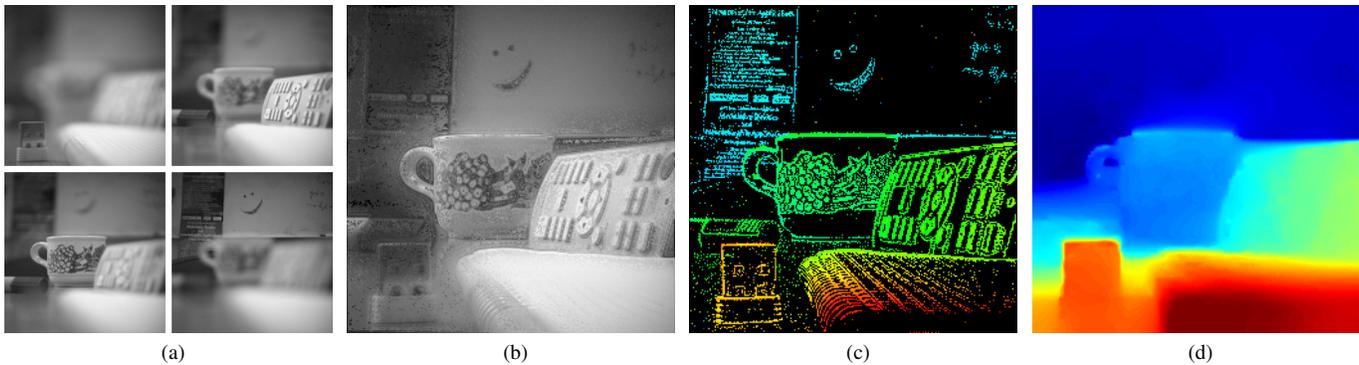
Fig. 7. An example scene imaged by our system: (a) shows four images extracted during a focal sweep illustrating the narrow DoF of our system; (b) shows an extended depth of field image (all points are in focus); (c) shows a depth map with 64 levels as captured by our system; (d) shows an inpainted, denoised and densified image as an example of post-processing that can be performed on images produced by our system.

in which every pixel is assigned a depth value. Hence, an interesting problem to tackle is to infer the depth value for missing data, i.e. for pixels which have no depth measurement because of the lack of texture or because of noise in our system. In our images, pixels with no depth value are shown in black. For instance in Figure 7, only about 15% of the pixels have a value. The problem of inferring missing pixel value in an image is known in the computer vision literature as inpainting.

In addition, one observes the presence of noise, whose sources are multiple. For instance, in the algorithm design, the criterion used to estimate focus is not really invariant to intensity: it could be that a light source that is out of focus still produces a strong LoG compared to an edge in a region of the image with poor contrast. Some other sources of noise come from the implementation on our focal plane processor, for instance, the analog computation of the criterion in SCAMP-5 introduces errors. Digital noise due to the flips of bits in the digital registers also happen. Hence, a model that inpaints our sparse images has to include an intrinsic denoising mechanism to avoid filling in the image with noise supporting wrong data.

Finally, a specificity of our images is that the depth levels are heavily quantized: the exact depth cannot be resolved within the range of the depth-of-field for a given distance as we outlined in II-B. A challenge is to infer a "continuous" depth value from the discrete depth levels obtained in our system. The continuous value should be both constrained to lie within the range prescribed by the level at which it was maximally in focus and its associated depth of field and to be close to the value of the neighbour pixels as a way to regularize the problem. We call this process – that aims to constrain the latent depth value to be continuous while it is based on discrete measurements – "densification".

In future work, we are interested in modeling jointly the three problems of inpainting, denoising, and densification for images produced by our system. Ultimately we also want to design an algorithm that can make use of local information, so that it could run on our focal plane processor along the depth imaging routing. The design of such an algorithm is out of the scope of this work but a preliminary result implementing a variational model to perform inpainting, denoising, and densification is presented in Figure 7d.

VI. CONCLUSION

We presented a system that can reconstruct depth from focus in real-time, using a vision sensor coupled with a high-speed tunable focus liquid lens. The $1.6W$ to $1.9W$ system records up to 128 depth levels, and at reduced depth resolution has been demonstrated to capture up to $150FPS$. This performance is made possible by a few key features of our setup. The vision chip collocates sensing and processing, which enables large internal bandwidth between a photosensor and processor array. The fast focal sweeps performed by the liquid lens reduce the complex problem of depth evaluation in an image to a set of comparatively simple focus measurements. Finally, the computation needed to evaluate and compare focus across images is formulated in a pixel-parallel algorithm suitable for the high-performance processor array.

ACKNOWLEDGMENT

REFERENCES

[1] A. G. Kashani, M. J. Olsen, C. E. Parrish, and N. Wilson, "A review of lidar radiometric processing: from ad hoc intensity correction to rigorous radiometric calibration," *Sensors*, vol. 15, no. 11, pp. 28 099–28 128, 2015.

[2] J. J. Leonard and H. F. Durrant-Whyte, "Directed sonar sensing for mobile robot navigation," *Springer Science & Business Media*, vol. 175, 2012.

[3] D. Fofi, T. Sliwa, and Y. Voisin, "A comparative survey on invisible structured light," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 90–98.

[4] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: From software to hardware," *International Journal of Optomechatronics*, vol. 2, no. 4, pp. 435–462, 2008.

[5] W. A. Simpson, "Optic flow and depth perception," *Spatial Vision*, vol. 7, no. 1, pp. 35–75, 1993.

TABLE I
SUMMARY OF TWO CONFIGURATIONS PROPOSED AND TESTED FOR OUR SYSTEM.

	Short-range configuration	Mid-range configuration
Array size	256×256	
Combined aperture of the system	16mm	
Number of depth levels	up to 128: 32 levels at 25 FPS under 10klx	
Power consumption	1.6W to 1.9W total 66% lens 34% vision chip	
Focal length of the objective-lens	13.5mm $f/1.2$	25mm $f/1.4$
Field of View	$51°$	$23°$
Depth of Field (resolution) ...	$\{0.01, 0.27, 11.0, 47.0, 225.0\}$cm	$\{0.1, 4.0, 16.5, 50.0, 400.0\}$cm
... at distances	$\{0.02, 0.25, 0.5, 1.0, 4.5\}$m	$\{0.1, 0.5, 1.0, 5.0, 8.0\}$m

[6] J. N. P. Martel, L. K. Muller, S. J. Carey, and P. Dudek, "High-speed depth from focus on a programmable vision chip using a focus tunable lens," in *Proc. of the IEEE Internat. Symp. on Circuits and Systems, ISCAS'17*, 2017.

[7] S. J. Carey, D. R. W. Barr, A. Lopich, and P. Dudek, "A 100,000 fps vision sensor with embedded 535 GOPS/W 256×256 SIMD processor array," in *Proc. of the VLSI Circuits Symp.'13*, 2013, pp. C182–C183.

[8] M. Blum, M. Baeler, C. Grätzel, and M. Aschwanden, "Compact optical design solutions using focus tunable lenses," in *Proc. of the SPIE Optical Design and Engineering IV*, vol. 8167, 2012.

[9] B. K. P. Horn, "Focusing," MIT Artificial Intelligence Laboratory, Tech. Rep., May 1968, memo 160.

[10] S. K. Nayar, "Shape from focus system," in *Proc. of Computer Vision and Pattern Recognition Conference, CVPR'92*, 1989, pp. 302–308.

[11] C. Zhou, D. Miau, and S. Nayar, "Focal sweep camera for space-time refocusing," Technical Report, Department of Computer Science, Columbia University, Tech. Rep., Nov. 2012.

[12] S.Kuthirummal, H. Nagahara, C.Zhou, and S. Nayar, "Flexible depth of field photography," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, 2010.

[13] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *Proc. of Computer Vision and Pattern Recognition Conference, CVPR'15*, 2015, pp. 3497–3506.

[14] J. N. P. Martel, L. K. Muller, S. J. Carey, and P. Dudek, "Parallel hdr tone mapping and auto-focus on a cellular processor array vision chip," in *Proc. of the IEEE Internat. Symp. on Circuits and Systems, ISCAS'16*, 2016.

[15] K. Boahen, "A throughput-on-demand address-event transmitter for neuromorphic chips," in *Proc. ARVlSI 99*, 1999, pp. 72–86.

Stephen Carey is a Research Fellow at the Microelectronics Lab at the School of Electrical and Electronics Engineering, The University of Manchester. His main interest are in the area of integrated circuit design in the development of novel sensor-processor systems.

Jonathan Müller is a Master student in Information Technology and Electrical Engineering at ETH Zurich. His main interests lie in the design of systems for event based visual sensing and processing, as well as parallel computing.

Julien Martel is a Ph.D candidate at the Institute of Neuroinformatics, University of Zurich and ETH Zurich. His main interest is in the design of novel algorithms and systems for visual sensing and processing.

Yulia Sandamirskaya is a Group Leader at the Institute of Neuroinformatics, University of Zurich and ETH Zurich. Her main interest is in the development of cognitive architectures for neuromorphic robots.
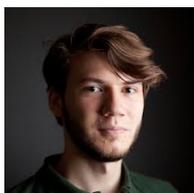
Lorenz Müller is a Post-doctoral researcher at the Institute of Neuroinformatics, University of Zurich and ETH Zurich. His main interest is in neurally inspired information processing.

Piotr Dudek is a Professor of Circuits and Systems in the School of Electrical and Electronic Engineering, The University of Manchester, leading the Microelectronics Design Lab. His research interest are in the area of integrated circuit design, especially vision sensors, cellular processor arrays, analogue and mixed-mode processing hardware, neuromorphic engineering and brain-inspired systems.