High-speed Depth from Focus on a Programmable Vision Chip Using a Focus Tunable Lens

Julien N. P. Martel *, Lorenz K. Müller *, Stephen J. Carey † and Piotr Dudek †
Email: jmartel@ini.ethz.ch lorenz@ini.ethz.ch stephen.carey@manchester.ac.uk p.dudek@manchester.ac.uk
* Institute of Neuroinformatics, University of Zurich / ETH Zurich, Zurich 8057, Switzerland
† School of Electrical & Electronic Engineering, The University of Manchester, Manchester M13 9PL, United Kingdom

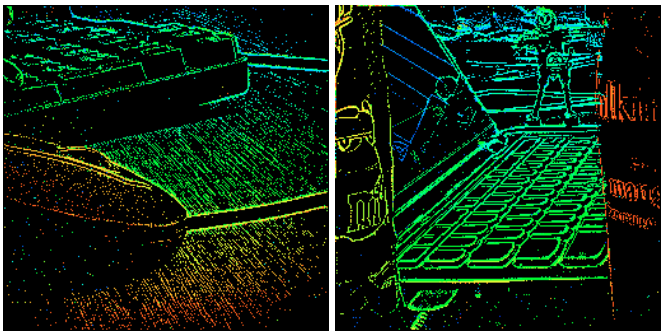
***Abstract*—In this paper, we present a 3D imaging system providing a semi-dense depth map, using a passive, low-power, compact, static, monocular camera. The demonstrated depth estimation system reconstructs 32 depth-levels in real-time at 25FPS drawing less than 1.9W of power. This is achieved by performing computation on an analog focal-plane processor that analyses frames captured through a vibrating liquid focus-tunable lens. The optical system provides shallow depth of focus images and fast sweeps of optical power, while the use of pixel-level processing removes the sensor-processor bandwidth limitations of depth-from-focus systems built using conventional imaging and processor technologies. All-in-focus images are also obtained.**

I. Introduction

The estimation of the three-dimensional structure of an environment is a challenging problem with many practical applications; among them are navigation, mapping or object segmentation. As suggested by the variety of available depth sensors, such as LIDARs, stereo cameras, structured-light, or optic-flow based devices, there exist a number of cues that can be used to estimate distances from an observer in a scene. Such systems, and the algorithms they run, vary widely in their computational complexity, form-factor, power-budget and thus in their domain of applicability: indoor, outdoor, on mobile platforms, depending on the light conditions or the required working-range for instance.

In this work we present a monocular depth imaging system, that does not require the active emission of energy into the scene, does not need the motion of the camera, and is computationally inexpensive. It is based on the limited depth of field created by "large aperture" optical systems. By changing the optical power of our system, objects at different depths in the imaged scene fall in and out of focus, i.e. they appear more or less sharp. We use a vision chip [**?**] (focal plane processor array) that we equip with a focus-tunable lens [**?**] configured to sweep focus back and forth at high-frequency. For each focus sweep, multiple images are taken and their sharpness is analyzed by the focal plane processor. Thereby, points in focus for an image taken at a given optical power – that is reached at a particular time during the sweep – can be associated with a depth value as we present in Section **??**. Such approach requires capturing and processing a large number of images to produce a single depth image frame, which limits the frame rate achievable when using conventional cameras. This limitation is alleviated in our system thanks to the computational capabilities of the focal-plane processor and the ultra-high bandwidth it has with pixel data; details about the system's implementation are given in Section **??**. The analysis of focus for each image is thus carried out on-chip without the need of an external readout. We show in Section **??** that we obtain sparse depth images with 32 levels at about 25FPS (where readout is currently the limiting factor) and show we can obtain up to 64 levels with a lower frame rate.

This work was funded by SNF grants 143947 and CRSII2_160756, EPSRC grant EP/M019284/1, and Royal Academy of Engineering Leverhulme Trust Senior Research Fellowship grant.



(a) (b)

Fig. 1. Two examples of 32-level sparse depth maps (recovered where focus can be estimated) as output by our system. The heat map encodes "near" distances in red and "far" in blue. **??** shows a mouse and keyboard on a wooden table **??** shows a cluttered scene.

II. Principles and algorithm in Depth from Focus

A. A basic model of focus

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 in [**?**]. 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 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 $\hat{d}_o + \hat{d}_i = \delta$. If one places an imaging sensor, orthogonal to the OA, at distance d_i, one images a point, and the system is said to be "in-focus for the object point at distance d_o". If the object is translated at $d'_o \neq d_o$, keeping

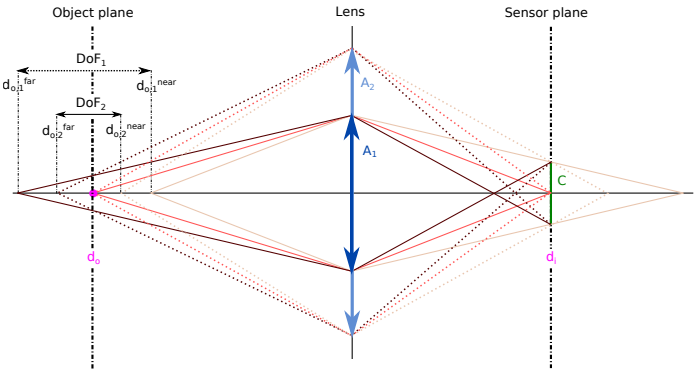


Fig. 2. An illustration of the depth of fields $DoF_{1/2}$, near and far limits of focus $d_o^{\text{near/far}}$, for two apertures $A_{1/2}$ using a convex-lens in front of a pixel defining the circle of confusion C, upper limit of resolution.

d_i fixed, then the object point projects onto the imager's plane as a circle that concentrates the same energy: the image point is blurred. Thereby, if an object forms a sharp image on a plane 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: its depth – its distance to the lens – is thus $\hat{d}_o = \delta - \hat{d}_i$.

One is generally not interested in solely imaging objects lying at a single depth but in real scenes containing objects that may be placed at different distances from the lens. Considering the simple model described previously, one has to vary one of the three parameters d_o, d_i or δ to bring in focus different object planes and get access to a range of depth values:

- *Varying d_o, the relative distance lens-object* can be achieved by "translating" the scene so that 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 [**?**].
- *Varying d_i, the relative distance lens-sensor* 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 [**?**] and [**?**].
- *Varying δ, the optical power* of the lens (or multi-lens equivalent) can be achieved using a mechanical system, and camera objectives equipped with motorised focus rings are widely available. Computer vision has traditionally addressed the problem of recovering Depth from Focus/Defocus offline, using such focal stacks (images taken at different focus) for static scenes [**?**]. However, if one aims at designing a high-speed depth imager, the slowness of such mechanical systems becomes an inconvenience.

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 up to a kilohertz.

B. Depth of field limitations

To better understand the design challenges and limitations of a system such as ours, 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. Also an imaging system does not present an infinite resolution but is ultimately limited by the size of its pixel, below which it is not possible to distinguish focus: This defines the circle of confusion C. With such modeling assumptions, one can observe that instead of a single point, a range of points in front of the lens now project below the discrimination limit we defined as our circle of confusion and therefore all fall in focus as illustrated in Figure **??**. The limits of the near-focus d_o^{near} and far-focus d_o^{far} can be derived, when focusing at d_o as:

$$d_o^{\text{near/far}} = \frac{d_o A f}{A f \pm C(d_o - f)} \tag{2}$$

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:

$$DoF = \frac{2 d_o A f C (d_o - f)}{A^2 f^2 - (d_o - f)^2 C^2} \tag{3}$$

By inserting $d_o = \frac{f\, d_i}{d_i - f}$ from Equation (**??**) in Equations (**??**) and (**??**) one can describe them only in terms of the intrinsic parameters of the system: d_i, f, C, and A. We use these formulas to recover, given a particular f setting of the focus-tunable lens, the range of depths corresponding to points in focus in the image.

Equation (**??**) yields the following important results: (a) as d_o increases the DoF becomes very large, since C is orders of magnitude smaller than $A\,f$, which means 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 **??** would yield an infinitely small DoF and infinitely high resolution.

C. An algorithm for depth from focus

As we hinted previously, our depth from focus algorithm takes advantage that for fixed lens-object distance d_o, and fixed sensor to lens distance d_i, the optical power of the lens δ can be changed to bring objects in focus.

Assume for now that we can sweep the optical power of the lens periodically through time $\delta : t \to \delta(t)$: all objects lying between the closest near limit of focus ($d_{o,\text{min}}^{\text{near}}$ associated to the highest reachable optical power δ_{max}) and the furthest limit of focus ($d_{o,\text{max}}^{\text{far}}$ corresponding to the lowest reachable optical power δ_{min}) will be brought in and out of focus during this focal sweep and will be sharp in the image for some time interval. Consequently, we can sample a focal sweep by taking, say N images and analyze their focus with a criterion that is meant to be maximal when points in the image are the sharpest.

We demonstrated in a previous work [**?**] that a Laplacian of Gaussian (LoG) is a convex criterion for focus and can be computed efficiently on the class of architectures we address. We then "search" for the sample index that, during the focal sweep maximizes this focus metric: this sample index relates to an optical power, and therefore to a particular range of depth using Equation (**??**). In practice, because focus is swept continuously and the criterion is chosen to be convex, a single pass keeping track of the maximum of focus is implemented. Besides, the comparison of this maximum to a user-defined threshold θ discards spurious noise by keeping only pixels

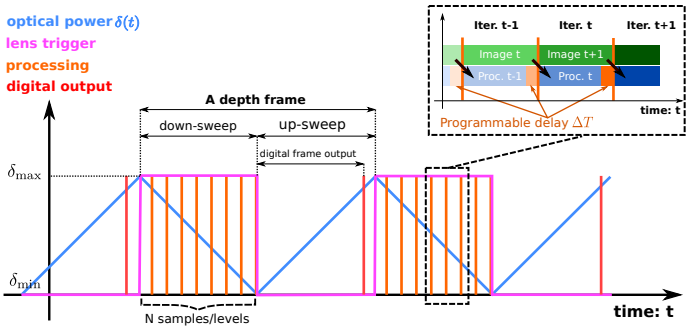


Fig. 3. A diagram illustrating timing in our depth from focus system and its pipelined imaging/processing operations.

in the scene where the focus metric reaches sufficiently high values. As a result of this algorithm, each sweep of focus, with N images produces a single depth frame with N depth levels. This algorithm is summarized in the following:

Require: R_1, R_2, R_{dig}, R_{im}: four register arrays
Require: R_{pix}: the image acquisition array (photosensors)
Require: N, ΔT: two parameters
$R_2 \leftarrow 0$ ▷ Stores the max focus metric value so far
for $n \in \{1, \dots N\}$ levels **do**
  $R_{\text{im}} \leftarrow R_{\text{pix}}$ ▷ Loads current image
  $R_{\text{pix}} \leftarrow 0$ ▷ Starts light integration for next image
  $R_1 \leftarrow f(R_{\text{im}})$ ▷ Compute focus metric
  if $R_1 > R_2$ **then** ▷ Compare focus to the max focus
    $R_2 \leftarrow R_1$ ▷ Store max focus
    $R_{\text{dig}} \leftarrow n$ ▷ Store iter. num.
  end if
  wait ΔT ▷ Control R_{pix} exposure
  $n \leftarrow n + 1$
end for
if $R_2 < \theta$ **then**
  $R_{\text{dig}} \leftarrow 0$ ▷ Discard pixels w/. low maxima
end if
trigger depth-frame readout of R_{dig}

Note that Equation (**??**) suggests an optimal way to sample the focus sweep. The "most-efficient" way to sample – so that the frame rate, and the number and distribution of depth levels are optimized – is to produce non-overlapping DoF. To do so, either the sampling procedure needs to be non-linear (cf. Section **??**) for a linear sweep, or vice-versa. We do not address this issue in the scope of this work.

III. Description of the System and Implementation

We use a pixel-parallel processor array vision chip [**?**] placed behind a fast focus-tunable lens [**?**] to perform the algorithm presented in Section **??**. The benefit of using such a device 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_{ana_out} is the readout time of an analogue, or 8-bit, image array) and $T_{\text{ana}\_out} \gg T_{\text{dig}\_out}$.

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 **??** and **??**. The tunable lens we use is a liquid-lens [**?**]. 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.
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 $[0.01, 3.5]\text{m}$ and $[0.1, 11]\text{m}$ 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 [**?**]. 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 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 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.

C. Implementation and configuration of the system

The focus-tunable lens is driven by a current generator mapping linearly a range of $[0, 191]\text{mA}$ to $[\delta_{\text{min}}, \delta_{\text{max}}] = [5, 10]\text{dpt}$ of optical power translating to $[-1.5, 3.5]\text{dpt}$ 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 **??**, 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 (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 get 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 $n \in \{1, \dots N\}$ occurs when imaging the $n + 1$

(a) (b) (c)

Fig. 4. An example scene imaged by our system **??** shows four images extracted during a focal sweep illustrating the narrow DoF of our system **??** shows an extended depth of field image (all points are in focus) **??** shows a depth map with 64 levels as captured by our system.

sample. Since processing is very fast, we set the delay ΔT to increase the exposure time to a useful range. Only two degrees of freedom exist since if one wants to maximally span the focal sweep (to get the maximal range of depth available) the following must hold:

$$\left(N \cdot (T_{\text{proc}} + \Delta T) = \frac{1}{2\nu}\right) \wedge \left(\lceil \log_2(N) \rceil \cdot T_{\text{dig\_out}} \leq \frac{1}{2\nu}\right) \quad (4)$$

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 Section **??** and $T_{\text{dig\_out}}$ a constant measuring the time to threshold and output a 1-bit plane digital register from the focal plane processor array. When the parameters ν, N and ΔT satisfy both Equations (**??**), 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.

IV. Results

In Figure **??** we show images as output from our depth from focus system, N=64 levels are mapped using a heat-map look-up table which indicates the depth R_{dig}; colors correspond to the index of the the level at which the the maximal focus was achieved. We can also record extended depth-of-field frames, as shown in Figure **??**, by storing the current pixel value when the maximal focus is reached. Using a $f = 25\text{mm}$ lens with a clear aperture of $A = 16\text{mm}$ at distances $d_o = \{0.1, 0.5, 1.0, 5.0\}\text{m}$ the corresponding depth of fields are $DoF = \{0.1, 4.0, 16.5, 50.0\}\text{cm}$. 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. We have demonstrated operation at 25FPS with N=32 levels. 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, $T_{\text{dig\_out}} = 8.4\text{ms}$ in contrast to a processing time of $T_{\text{proc}} = 56\mu\text{s}$ 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. Also, 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, 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 then largely benefit from the use of a non-linear sampling scheme as suggested in Section **??**. To speed up the digital output, a sparse Address Event Representation (AER) 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.

V. 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 large internal bandwidth between the photosensor and processor arrays, and high performance of the processor array, allow the implementation of the entire depth-frame reconstruction algorithm on-chip, in real-time, without the need of any external processing.