Compressive extended depth of field using image space coding

Patrick Llull, Xin Yuan, Xuejun Liao, Jianbo Yang, Lawrence Carin, Guillermo Sapiro, and David J. Brady*

Fitzpatrick Institute for Photonics, Department of Electrical and Computer Engineering
Duke University, 129 Hudson Hall, Durham, North Carolina 27708, USA
Corresponding author: dbrady@duke.edu

Abstract: We present a prototype imaging system that utilizes focal and temporal image space modulation to compress high-frequency, volumetric spatiotemporal information into low-framerate, planar images. 24-frame video sequences and extended depth of field reconstructions from single compressive measurements are reported. OCIS codes: (110.1120) Apertures; (110.1758) Computational imaging

1. Introduction

Aside from increasing system F/#, single-aperture extended depth of field (EDOF) techniques utilize focal stacks or sweeps [1], depth-from-focus and defocus algorithms [2], pupil coding [3], or radiance field coding [4] to estimate an all-in-focus image from one or multiple 2-dimensional (2D) images. [5] showed that temporal resolution is reduced when capturing a focal stack, spatial resolution decreases when coding the radiance field, and modulation transfer efficiency, SNR, and chromatic performance are reduced when coding the pupil.

Here, we propose image-space coding [5] to estimate a sequence of video frames and an EDOF image from a single 2D, coded measurement. In general, image-space coding techniques exploit the correlation of ambient features by modulating the impinging datastream in the compressed dimension. For our application, we may extrapolate the temporal or focal bandwidth, based on two inversion strategies described within.

We augment the temporal superresolution framework established in [6] with liquid-lens-based focal modulation to enable 4D volumetric and temporal multiplexed sampling and reconstruction from 2D measurements. Liquid lenses have been used in the focal stacking and EDOF literature [7] to scan the object volume rapidly and accurately.

2. Coding Strategy and System Model

Considering a single transverse spatial dimension ((x,y) → x), let (x,z,t) respectively denote the three orthogonal spatial coordinates and the temporal coordinate. Additionally, let object-and image-space coordinates be respectively designated with unprimed and primed variables. Given an integration time Δ, two-dimensional space-time compressive images g(x′,t′) are formed on the detector, with t′ < t. The digital data used to represent the scene consists of discrete samples of the continuous transformation [8]

\[
g(x',t') = \int \int \int f(x,z,t) T(x-s(t)) \text{rect} \left( \frac{x-x'}{\Delta} \right) \text{rect} \left( \frac{t-t'}{\Delta} \right) dx dt dz,
\]

where \( f(x,z,t) \in \mathbb{R}^4 \) is the analog volumetric video to be estimated and \( T(x-s(t)) \) represents a random binary transmission pattern that translates with linear transverse motion parameterized by \( s(t) = vt \), where \( v = 12 \) pixels per integration time in our experiments. The spatial and temporal pixel sampling functions are respectively denoted \( \text{rect} \left( \frac{x-x'}{\Delta} \right) \) and \( \text{rect} \left( \frac{t-t'}{\Delta} \right) \). The coded aperture is used to modulate the space-time volume and alias high temporal frequencies into the passband of the detector sampling functions [6,8].

The optical datastream \( f \) is represented as a convolution: \( f = c * b \). Here, \( c(x,z,t) \) denotes an instantaneous all-in-focus representation of the scene and \( b(x,z(t)) \) is a local time-and depth-varying blur kernel that may be approximated as a Gaussian with standard deviation \( \sigma_b \). After this convolution, the modulated focal volume is projected onto the time-varying coded aperture \( T(x-s(t)) \) prior to integration at the detector plane to form the images \( g \) [6].

To obtain an \( N_F \)-frame global estimate of the optical datastream as in [6], we invert the vector system \( g = Tf \), where \( g, T, f \) are the vectorized measurement, the transmission matrix, and the desired datacube, respectively. Here, \( N_F = 24 \); thus, we approximate the code motion with 24 steps as in [6]. We use the GAP algorithm [9] for these short video sequence reconstructions. Using focal modulation with this reconstruction method has been shown to enable compressive focal stack capture [11].
To reconstruct an accurate EDOF image, we model the image formation at the patch level since Eq. 1 represents a shift-variant transformation. Let the $l^{th}$, $P 	imes P$ patch of the image be denoted $g_{(l)}$. Since we do not know the patch-level object’s location a-priori, we must choose from $N_z$ candidate depths $z$ to model the patch-level blur kernel standard deviations $\sigma_{b,(l)}$. A defocus wavefront error of $w_{020} \lambda$ waves equates to a focus error of $2 \lambda F/# \lambda w_{020}$, which results in a maximum transverse ray error of $2 \lambda F/# \lambda w_{020} [10]$. Dividing the second expression by the first gives us an angle that may be multiplied by the focal difference to yield

$$\sigma_{b,(l)}(z_f, z_i, t) = \left(1 - \frac{1}{e}\right) \frac{|z_f - z_i|}{4(F/#)}$$

where $F(t)$ is the time-varying focal length of the compound liquid lens. The in-focus object distance as a function of time is $z_f$; the image distance $z_i$ is the image distance of the current candidate distance $z$.

To capture every part of the scene in sharp focus at some time during $\Delta t$, we employ a hardware setup similar to that used in [6], with the addition of an EL-10-30-VIS-HR liquid lens (Optotune) spaced 7cm apart from a 12mm fixed lens (Lensation). The detector (AVT Pike) has 7.4μm pixels; the coded aperture has 14.8μm features. A time-varying, 15Hz triangle wave voltage signal from a function generator simultaneously triggers a piezoelectric stage to translate the coded aperture a stroke of 88μm at 15Hz, the liquid lens to change the system focal length between 11.2 and 9.7 mm at 15Hz, and the camera to capture images at 30fps.

The patch-based model at an axial plane $z$ is given by $g_{(l),z} = T_{(l)} \cdot B_{(l),z} \cdot c_{(l)}$, where $B_{(l),z} \in \mathbb{R}^{P^2 \times 4P^2}$ models the local blur function candidate $b_{(l)}(x,y,z(t))$ for the $l^{th}$ patch. These matrices are extended by a factor of 4 in the horizontal dimension to account for convolution boundary effects. In our experiments, $\sigma_b$ has a maximum value of $\sim$ 8 pixels (59.2μm); we use $P = 16$.

Since we model the liquid lens’ focal modulation as a multiplexing operation of $N_z$ discrete object depths onto the detector, the blur matrix $B_{(l),z}$ must be extended along the vertical dimension to account for the series of blur kernels imparted on the focal volume:

$$g_{(l)} = B_{(l)} \cdot c_{(l)} = T_{(l)} \begin{bmatrix} B_{(l),z_1} \\ B_{(l),z_2} \\ \vdots \\ B_{(l),z_{N_z}} \end{bmatrix} \cdot c_{(l)} \quad (3)$$

where the additional subscript $z$ denotes the candidate object distance that generates the blur kernel (Eq. 2).

Eq. 3 can be rewritten to support $N_z$ candidate blur matrices (the current candidate denoted with a superscript $s$) as $g_{(l)} = T_{(l)} B_{(l)}^{(s)} \cdot c_{(l)}$, where each of the $N_z$ blur matrix candidates $B_{(l)}^{(s)}$ are truncated matrix representations of Eq. 2.

We estimate the clear image $\hat{c}_{(l)}$ by solving

$$\hat{c}_{(l)} = \arg \min_{c_{(l)}} \| g_{(l)} - T_{(l)} B_{(l)}^{(s)} c_{(l)} \|_2$$

where $c_{(l)}$ is the reconstructed, candidate EDOF image of the $s^{th}$ candidate based on the current blur matrix $B_{(l)}^{(s)}$. The EDOF estimate consists of all the patches $\hat{c}_{(l)} \forall l \rightarrow \hat{c}$ reconstructed this way.
3. Experimental Results

To test the capabilities of the prototype camera, we obtain EDOF images of static (Fig. 1(a)) and dynamic (Fig. 2(a)) 3D scenes. In the case of zero voltage supplied to the liquid lens, the camera is focused approximately on the farthest objects (the birds in Fig. 1(a) and the resolution target in Fig. 2(a)).

![Figure 2: Dynamic volumetric scene (a) and its reconstructions. (b) Two frames of a dynamic volumetric scene from an uncoded imager. (c) Two [Equivalent] measurements from the prototype compressive camera. (d) The two reconstructed all-in-focus images from the EDOF camera. (e,f) close-ups of (b,d). (g) Four frames of the global GAP reconstruction for the datastream.](image)

The static scene reconstruction using the CACTI framework (Fig. 1(d)) of the global GAP algorithm [9] shows the birds nearly in focus and the lego significantly out of focus. The EDOF solution of Eq. 4 renders the reconstructed lego and birds nearly as sharp as in the uncoded case (compare (e) with the lego in (b) and the birds in (c). We use a local version of GAP to estimate the $c_i^{(s)}$ prior to solving Eq. 4, which yields the true in-focus image $c_i^{(l)}$ at each patch.

Reconstructions of the dynamic scene shown in Fig. 2(a) reveal the increased temporal resolution of the EDOF camera. The uncoded camera shows the motion-blurred ball and out-of-focus newspaper. From the compressive measurements (Fig. 2(c)) we obtain EDOF images (Fig. 2(d)). Note that, with the dynamic scene, there are reconstruction artifacts near the smiley face attributed to unknown motion blur. We use the global GAP algorithm to obtain a temporal video sequence (Fig. 2(g)) to superresolve the motion-blurred smiley face.

4. Conclusion

We have demonstrated an imaging framework capable of rendering a volumetric scene in focus or a video sequence from a single compressive measurement. Future work will aim to reconstruct full EDOF video sequences from a single snapshot.

References and links