MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Towards Motion-Aware Light Field Video for Dynamic Scenes

Tambe, S.; Veeraraghavan, A.; Agrawal, A.

TR2013-111 December 2013

Abstract

Current Light Field (LF) cameras offer fixed resolution in space, time and angle which is decided a-priori and is independent of the scene. These cameras either trade-off spatial resolution to capture single-shot LF [20, 27, 12] or tradeoff temporal resolution by assuming a static scene to capture high spatial resolution LF [18, 3]. Thus, capturing high spatial resolution LF video for dynamic scenes remains an open and challenging problem. We present the concept, design and implementation of a LF video camera that allows capturing high resolution LF video. The spatial, angular and temporal resolution are not fixed a-priori and we exploit the scene-specific redundancy in space, time and angle. Our reconstruction is motion-aware and offers a continuum of resolution tradeoff with increasing motion in the scene. The key idea is (a)to design efficient multiplexing matrices that allow resolution tradeoffs, (b) use dictionary learning and sparse representations for robust reconstruction, and (c) perform local motion-aware adaptive reconstruction. We perform extensive analysis and characterize the performance of our motion-aware reconstruction algorithm. We show realistic simulations using a graphics simulator as well as real results using a LCoS based programmable camera. We demonstrate novel results such as high resolution digital refocusing for dynamic moving objects.

IEEE International Conference on Computer Vision (ICCV)

Towards Motion-Aware Light Field Video for Dynamic Scenes

Salil Tambe, Ashok Veeraraghavan
Rice University
6100 Main St, Houston, TX 77005
[sst5, vashok] @rice.edu

Amit Agrawal
Mitsubishi Electric Research Labs (MERL)
201 Broadway, Cambridge, MA 02139
agrawal@merl.com

Abstract

Current Light Field (LF) cameras offer fixed resolution in space, time and angle which is decided a-priori and is independent of the scene. These cameras either trade-off spatial resolution to capture single-shot LF [20, 27, 12] or tradeoff temporal resolution by assuming a static scene to capture high spatial resolution LF [18, 3]. Thus, capturing high spatial resolution LF video for dynamic scenes remains an open and challenging problem.

We present the concept, design and implementation of a LF video camera that allows capturing high resolution LF video. The spatial, angular and temporal resolution are not fixed a-priori and we exploit the scene-specific redundancy in space, time and angle. Our reconstruction is motion-aware and offers a continuum of resolution tradeoff with increasing motion in the scene. The key idea is (a) to design efficient multiplexing matrices that allow resolution tradeoffs, (b) use dictionary learning and sparse representations for robust reconstruction, and (c) perform local motion-aware adaptive reconstruction.

We perform extensive analysis and characterize the performance of our motion-aware reconstruction algorithm. We show realistic simulations using a graphics simulator as well as real results using a LCoS based programmable camera. We demonstrate novel results such as high resolution digital refocusing for dynamic moving objects.

1. Introduction

Traditionally cameras have required photographers to make trade-offs in terms of depth of field (DOF), dynamic range, shutter speed and ISO during the capture itself. With the advent of computational photography, such decisions are being shifted to post-processing. This paradigm enables more user control over the captured photo. For example, LF cameras such as Lytro [20], Raytrix [27] etc. allow digital refocusing. However, even these computational cameras are similar to conventional cameras in forcing a-priori choice in space, time and angle resolution.

For example, a single-shot LF camera offers tradeoff between spatial and angular resolution and captures a low spatial resolution LF. The Lytro camera uses a 11 megapixel sensor but can acquire LF at $\approx 300 \times 300$ pixels spatial resolution. The tradeoff is *fixed* and is *scene independent*. This is a significant limitation of single-shot LF cameras. Another approach to capture LF video is to use multiple video cameras (*e.g.* ProFusion [26]), which is expensive and requires very high bandwidth. In addition, it requires accurate geometric and photometric calibration between the cameras. To enable high resolution LF[1], Liang *et al.* [18] captured several multiplexed coded aperture images and demultiplexed them. However, this requires the scene to be static, thereby trading off temporal resolution for high spatial resolution. Thus, it is clear that there exists a resolution barrier (Figure 1) for capturing high resolution LF video. Recently, a series of approaches such as Plenoptic 2.0 camera [12] and LF super-resolution techniques [4] have begun breaking this resolution barrier.

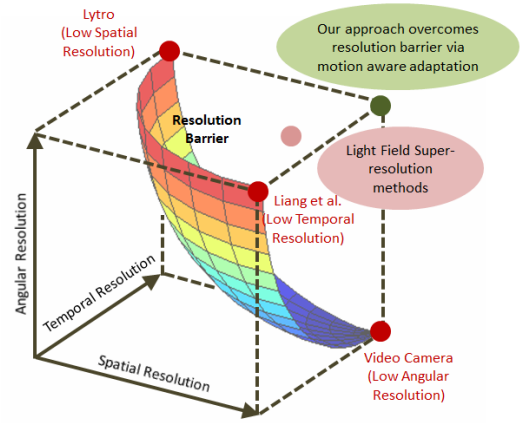

Figure 1. Current light field capture designs offer fixed, a-priori, and scene independent space-time-angle resolution. They are unable to cross the resolution barrier required for capturing high spatial resolution LF video. LF super-resolution techniques have recently begun breaking this resolution barrier. Our approach overcomes this barrier via motion-aware adaptive reconstruction using a programmable aperture camera.

[1]For the rest of the paper, high resolution LF refers to obtaining full spatial sensor resolution in LF reconstruction.

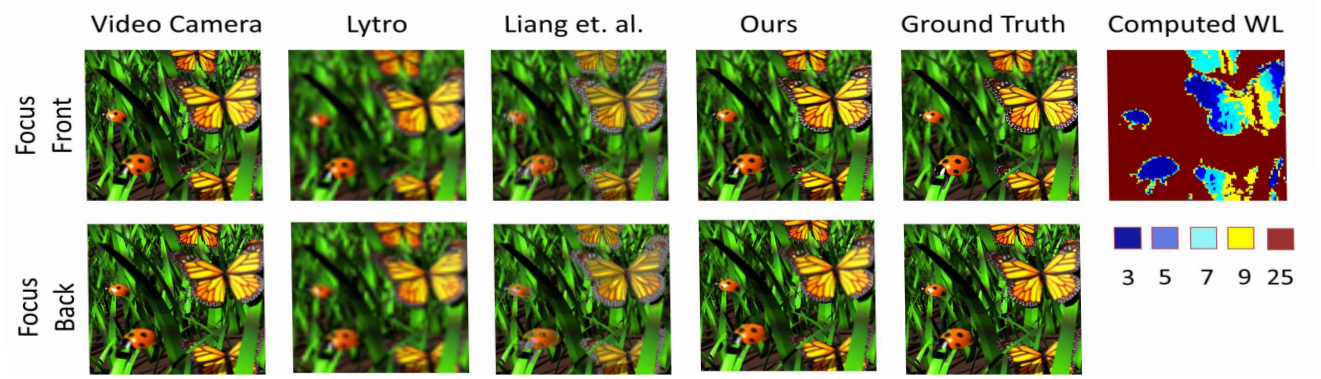

Figure 2. Simulated scene with moving butterfly and beetle and static grass. Notice the low spatial resolution refocusing of Lytro, as well as artifacts on moving objects for Liang *et al*. [18]. A standard video camera can either focus in front or back, but cannot achieve digital refocusing. Our approach allows capturing high spatial resolution LF for dynamic scenes. The final image on the right shows the distinct window lengths computed using our motion aware algorithm for this scene

In this paper, we take a step towards capturing high resolution LF video and present a computational approach for overcoming the resolution barrier using programmable aperture imaging. Our key concept is to overcome the fixed, a-priori and scene independent resolution trade-offs offered by previous cameras. Conceptually, we use several coded aperture patterns (one per time frame), which would allow reconstructing a high resolution LF if the scene was static. We repeat the patterns, and handle dynamic scenes by a *motion-aware* reconstruction; for each pixel, the number of frames used for reconstructing LF depend on its motion. While previous approaches have used Hadamard multiplexing for designing the codes [18], we learn them using dictionary learning (DL) and sparse representations. Thus, our design is a synergy between near-optimal patterns used for multiplexing and reconstruction algorithm.

Figures 2 and 3 show a motivating example of a dynamic scene with static grass and moving butterfly and beetle. Single-shot LF cameras such as Lytro lose significant spatial resolution. Liang *et al*. [18] can recover high spatial resolution LF but only on the static parts (grass) and show artifacts on moving objects (butterfly/beetle). Our approach provides high resolution LF for both moving and static scene parts.

1.1. Contributions

- We present the concept, design and implementation of a LF video camera and reconstruction algorithm that allows capturing high resolution LF video by analysing the spatial, temporal and angular resolution trade-offs.
- We propose a dictionary learning and sparse representation based algorithm for full resolution LF reconstruction and show how to adapt the algorithm to object/scene motion. We also show how to optimize the programmable aperture patterns using the learned dictionary.

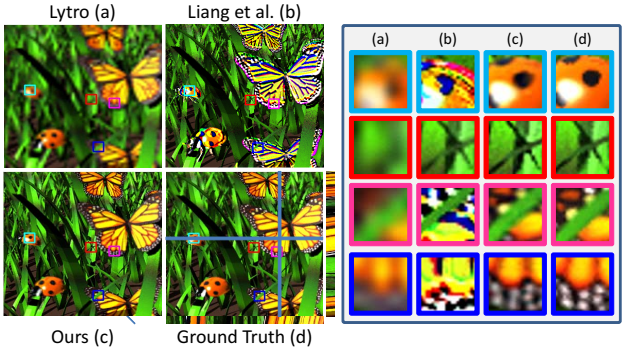

Figure 3. (Left) One of the angular LF 'view' using different approaches. Notice the low spatial resolution of Lytro, as well as artifacts on moving objects for Liang *et al*. [18]. The horizontal and vertical LF disparity is indicated for a row and column using ground truth. (Right) Zoom in of four regions indicated by different color outlines. Note that our approach results in high resolution LF information without any artifacts.

2. Related Work

Camera Arrays: One approach to capture LF video is to use a camera array [16, 17, 31]. Such approaches are hardware intensive, costly and require extensive bandwidth, storage and power consumption. As discussed, they require accurate geometric and radiometric calibration between different cameras.

LF capture: Existing LF cameras can be divided into two main categories: (a) single shot [24, 13, 30, 20, 27, 12, 22], and (b) multiple shot [18, 23, 3]. Most single shot light field cameras multiplex the 4-D LF onto the 2D sensor, losing spatial resolution to capture the angular information in the LF. Such cameras employ either a lenslet array close to the sensor [24, 12], a mask close to the sensor [30] or an array of lens/prism outside the main lens [13]. Recently, [22] extended the mask based method of [30] to exploit sparse representations in order to recover full resolution LF. Our method is similar in spirit but works to recover the loss

Method	Number of Cameras	Number of Images	Max. Spatial Resolution	LF Temporal Resolution	Hardware Design	Resolution Loss	Cost/ Bandwidth
Single-Shot [24, 30]	1	1	$\frac{N}{M} \times \frac{N}{M}$	Per frame	Mask/Lenslets close to sensor	Spatial	Low
Liang *et al*. [18]	1	M^2	$N \times N$	Per M^2 frames	Programmable Aperture	Temporal	Low
Camera Array	M^2	M^2	$N \times N$	Per frame	Multiple cameras	None	High
Ours	1	Scene Dependent	$N \times N$	Per frame	Programmable Aperture	Scene Dependent	Low

Figure 4. Comparison of different approaches to capture a $M \times M$ angular resolution LF using $N \times N$ spatial resolution sensor. Our approach offers a scene dependent resolution loss in space, time and angle. For static pixels, we utilize M^2 frames to get full resolution LF. As the pixel motion increases, less number of frames are utilized to reconstruct LF in a motion-aware fashion.

of temporal resolution in [18].

LF Super-resolution and Plenoptic2.0**:** Plenoptic cameras suffer from low spatial resolution. Recently, several LF super-resolution algorithms have been proposed to recover the lost resolution [12, 4]. The Plenoptic2.0 camera [12] recovers the lost resolution by placing the microlens array at a different location compared to the original design [24]. Similarly, the Raytrix camera [27] uses a microlens array with lenses of different focal length to improve spatial resolution. Thus, improving the spatial resolution of LF cameras is an active area of research.

Programmable Aperture Imaging: Programmable aperture imaging [18] allows capturing light fields at the spatial resolution of the sensor. In principle, each coded aperture can be a pin-hole placed at a different location in the aperture. A set of M^2 images are required to achieve an angular resolution of $M \times M$. To improve light efficiency, [18] use Hadamard multiplexed patterns. However, temporal resolution is sacrificed to achieve higher spatial resolution in LF. Recently, Babacan *et al*. [3] showed how to reduce the number of captured images by employing a Bayesian approach and a total variation based prior. Our approach is similar in spirit, but differ in following ways. Firstly, we learn a sparse basis dictionary from real LF data and use it along with the sparse reconstruction framework. Secondly, unlike [3], we adapt our reconstruction algorithm to the local motion of the scene, thereby preserving both motion and disparity information. Finally, we also search for near-optimal aperture codes so as to improve the reconstruction performance.

Resolution Tradeoffs: The *reinterpretable imager* by Agrawal *et al*. [1] has shown resolution tradeoffs in a single image capture. The results in [1] are on stop-and-go dynamic scenes and continuous motion cannot be handled. In contrast, our camera is a video LF camera running at 25 fps and can handle smooth motions. Agrawal *et al*. [1] require moving a slit/pinhole in the aperture and a static mask close to the sensor. Our design is simpler using only a dynamic coded aperture. More importantly, [1] can only achieve 1-D parallax information for dynamic scenes, while our approach enables parallax in both dimensions.

Coded Aperture: Coded aperture imaging has been widely used in astronomy [28] to overcome the limitations imposed by a pinhole camera. The concept of placing a coded mask close to the sensor for LF capture was proposed by [30]. Coded masks have also been used for estimating scene depth from single image [15],and for compressive LF [22] and video acquisition [21].

Compressive Sensing (CS): CS achieves below Nyquist rate sampling while enabling recovery of signals that admit a sparse representation in some basis [6, 8]. CS has been shown useful for light transport capture [25] and even LF capture [3]. However, these techniques still assume scene to be static for the duration of captured images and cannot handle moving objects. Recently, adaptive methods for enhancing the fidelity of CS video reconstruction have also been proposed [32].

3. Programmable Light Field Acquisition

Consider the two-plane parameterizations of the light-field $LF(u, v, s, t)$, where (u, v) represents co-ordinates on the aperture plane and (s, t) represents co-ordinates on the sensor plane. Let us assume that the aperture can be divided into $M \times M$ sub-apertures. The light field can then be obtained by capturing M^2 distinct photos, with only one of the M^2 distinct sub-apertures open in each of the images. The spatial resolution of the captured LF is determined by the sensor resolution, while the angular resolution is determined by the number of sub-apertures (and is equal to the number of images acquired). This results in a high spatial resolution, but the scene is assumed to be static for the entire capture duration (of M^2 photos). Any motion of scene elements during the acquisition time results in significant reconstruction artifacts (see Figures 2 and 3).

3.1. Dynamic Light Fields

Now we describe our approach to handle dynamic scenes in LF acquisition. Conceptually, we also use several coded aperture patterns (one per frame), which allows reconstructing a high resolution LF if the scene was static. However, our approach differs from [18] in following ways.

Firstly, we learn optimized dictionaries and coded aperture patterns that along with sparsity regularized reconstruction algorithms allow for the recovery of light fields from as few as three captured frames. Secondly, we repeat the patterns and using *overlapping* windows of temporal frames to

perform reconstruction. This ensures that our approach can handle *continuous* motion. Finally, we use a *motion-aware* reconstruction; for each patch, the number of frames used for reconstructing LF depend on its motion. Intuitively, if the scene is static, one should use more images for reconstruction. Our motion-aware reconstruction automatically chooses the best window length for each patch. We learn the mapping between the motion and window-length *off-line*. At run time, we compute the optical flow in the scene, and decide the window length using the above mapping.

4. Motion-Aware Adaptation for Light Field

We now describe the two key algorithmic aspects of our approach: (a) compressive LF sensing, and (b) motion-aware reconstruction.

4.1. Compressive LF Sensing

Consider a programmable LF camera with spatial resolution $N \times N$ pixels and angular resolution $M \times M$. Let $c_t(u, v)$ denote the coded aperture used at frame t. The captured frame I_t can be written as

$$I_t(s,t) = \sum_{(u,v)} c_t(u,v) \times LF(u,v,s,t). \tag{1}$$

The summation over u, v is representative of the integral of 4-D LF along angular dimensions to form a 2-D image. Since the above modulation is only in the angular dimensions of the LF via the mask, the equation is valid for any LF patch. Let us consider a $P \times P \times M \times M$ patch of the LF and vectorize it into a vector $\mathbf{x}_t$ of length P^2M^2. Let us also vectorize the corresponding captured image patch of size $P \times P$ pixels into a vector $\mathbf{y}_t$ of length P^2. Each captured image results in a linear set of equations given by

$$\mathbf{y}_t = C_t \mathbf{x}_t, \tag{2}$$

where C_t is a $P^2 \times P^2M^2$ matrix that encodes the aperture code used at time frame t. Taking F consecutive frames (with different codes), and assuming that the patch remains stationary during these F frames, the above linear system can be concatenated as

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_F \end{bmatrix} = \begin{bmatrix} C_1 \\ \vdots \\ C_F \end{bmatrix} \mathbf{x}_t \tag{3}$$
$$\mathbf{y}_{1:F} = C_{1:F}\mathbf{x}_t,$$

where $C_{1:F}$ is a $FP^2 \times P^2M^2$ matrix.

In [18], $F = M^2$ frames were acquired so as to be able to perform linear inversion for *every* pixel. In contrast, our motion-aware algorithm chooses a different F for each patch depending on its motion. However, when $F < M^2$ (required to handle dynamic scenes), the above system of linear equations is under-determined. Therefore, we need to enforce regularization conditions on the LF patch $\mathbf{x}_t$ in order to invert the linear system.

4.2. Dictionary Learning based Prior

We assume that the LF patches $\mathbf{x}_t$ are *sparse* in an over-complete dictionary. Over-complete dictionaries that are learned from data have been used successfully for a wide variety of imaging applications including image denoising [11], video recovery [14], deblurring [7], super-resolution [7] and image based classification [33].

It is important to learn a good dictionary that can faithfully represent the light fields we intend to capture. The quality of a dictionary is decided by its ability to reliably reconstruct light fields with varying amounts of (a) disparity, (b) texture, and (c) occlusion relationships. For learning the dictionary, we render light fields in a graphics rendering engine (`Povray`) with varying texture, disparity and occlusions (*e.g.* foreground object obstructing a background object). We first extract all $8 \times 8 \times 5 \times 5$ patches from the set of rendered light fields. We then use the K-SVD [2] algorithm to learn $K = 10000$ atoms from the 0.5 million extracted patches.

As the patch size increases, the learned dictionary can better capture the disparity dependent redundancies in the LF, thereby improving the reconstruction performance. However, a larger patch size also leads to a bigger dictionary and slower reconstruction. Thus, we restrict ourselves to using a patch size of $8 \times 8 \times 5 \times 5$.

4.3. Reconstruction Algorithm

Let the $P^2M^2 \times K$ matrix D represent the learned LF dictionary containing K atoms. A LF patch $\mathbf{x}$ can be written as $\mathbf{x} = \mathbf{Ds}$, where $\mathbf{s}$ is a sparse vector representing the coefficients. We use the learned dictionary D as a sparse regularizer for the under-determined system of linear equations and solve the following optimization problem:

$$\mathbf{P}_1 : \quad \min_{\mathbf{x}_t} ||\mathbf{y}_{1:F} - C_{1:F}\mathbf{x}_t||_2 + \lambda||\mathbf{s}||_0 \quad \text{subject to} \quad \mathbf{x}_t = \mathbf{Ds}, \tag{4}$$

where λ is a constant. We use orthogonal matching pursuit algorithm [29] to solve the optimization problem $\mathbf{P}_1$ for each patch in the LF. We use 4 adjacent overlapping set of patches shifted by 2 pixels, obtain individual reconstructions for each of the overlapping patches and average the reconstructions to produce the final result. The quality of the reconstructed LF depends upon the number of frames (or the window length) of the reconstruction.

4.4. Motion-Aware Reconstruction

When the scene is static, increasing the number of frames improves performance because each additional frame provides additional information for linear system inversion (de-multiplexing). Figure 5 (left) shows the reconstruction PSNR as a function of the number of frames F (window length) used for reconstruction when the scene remains static (red plot). Note that the reconstruction PSNR is greater than 20 dB even when only 3 frames are used.

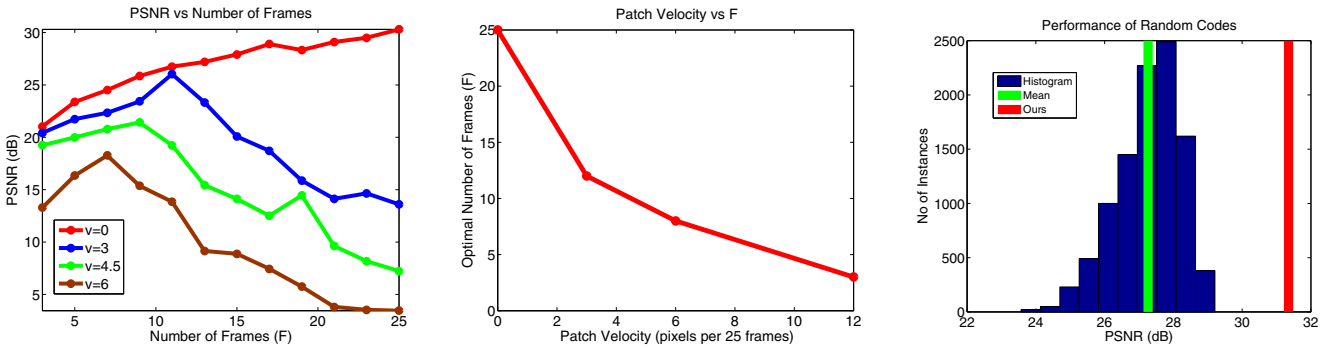

Figure 5. (Left) Plot of PSNR of reconstructed LF with number of frames F used with increasing object velocity. As the velocity increases, PSNR peaks and then degrades. (Middle) Plot showing the optimal number of frames used in motion-aware reconstruction as a function of the average patch velocity. (Right) Histogram of number of codes with PSNR for $10,000$ randomly generated codes. Our optimized code (red) performs 4.08 dB better compared to the histogram average.

However, for moving objects, increasing the number of frames also introduces modelling error in the linear system. Figure 5 (left) shows the reconstruction PSNR for different object velocities (3, 4.5 and 6 pixels per 25 frames). Notice that the reconstruction performance varies as a function of both object velocity and the window length(WL) used in reconstruction. In particular, the PSNR peaks for some WL and then decreases. Figure 5 (middle) shows that the optimal window length (which provides maximum PSNR) is an decreasing function of the velocity. We use this relationship between velocity and optimal window length in order to decide the number of frames used in reconstruction on a patch-wise basis.

Patch Velocity Estimation: Optical flow based approaches are common for estimating pixel velocity in video sequences. However, traditional flow estimation algorithms assume the same viewpoint for all the frames. Since each of our input frames is obtained using a different code in the aperture, it results in slight shift in viewpoint, leading to depth dependent disparity shifts between adjacent frames. We found that this results in slightly incorrect optical flow. We use the state-of-art optical flow technique by Liu [19] and mitigate this effect as follows. Firstly, while selecting aperture codes, we ensure that the average disparity between adjacent aperture codes are minimized. This helps in minimizing the viewpoint shift between the frames. Secondly, we blur the captured images before computing the optical flow. The estimated optical flow is used to choose the appropriate WL. The last image in fig.2 shows the computed WLs for different patches in the butterfly scene. Notice that the outer edges of the butterfly uses fewer measurements (3) for reconstruction(has a smaller WL) as compared to the body of the butterfly which uses a WL of 9 since the outer edges of the wing have more motion.

4.5. Optimizing Aperture Codes

Now we discuss how to optimize aperture codes to improve the SNR. The problem of finding a set of $M \times M$ optimal codes (equivalently, the mixing matrix C) for a given dictionary D is non-trivial. Approaches such as [9] that simultaneously learn the dictionary and sensing matrix do not take into account the inherent constraints imposed by the hardware, such as code being binary and non-negative (due to LCoS). Compressive sensing methods [5] utilize the Restricted Isometry Principle (RIP) property to design mixing matrix considering the *worst-case* performance. Instead, we choose to optimize over the *average* performance.

Let $E = CD$ be the effective multiplexing matrix. Similar to [10], we minimize the average mutual coherence between elements of the effective mixing matrix E to find the set of codes. This is defined as the average of all normalised inner products over columns in E. However, this optimization problem is non-convex. Therefore, we search over a million randomly generated binary codes and choose the one the minimizes the average mutual coherence. Figure 6 shows the 25 coded aperture masks that were found using this approach.

To demonstrate that our optimized patterns indeed improve performance, we compare it with the performance of $10,000$ randomly generated aperture codes. Figure 5 (right) shows the histogram of the number of randomly generated codes with PSNR. The performance of our optimized code (red line) is better by 4.08 dB compared to the mean performance of randomly generated codes.

5. Prototype

Our prototype system for motion-aware LF video capture uses a Liquid Crystal on Silica (LCoS) modulator as the spatial light modulator (SLM). Figure 6 shows a graphic illustration of the light path in our prototype and a photo of the actual hardware. The system uses a 1024×1280 pixel LCoS reflective SLM (SXGA-3DM) developed by Forth Dimension Displays. It can be controlled at a maximum frame rate of $\approx$ 5000 fps. We follow the optical design in [23]. The LCoS modulator is at the effective aperture plane of the imaging system. While the size of LCoS is 17.4×14.7 mm, we only use the central 10×10 mm area

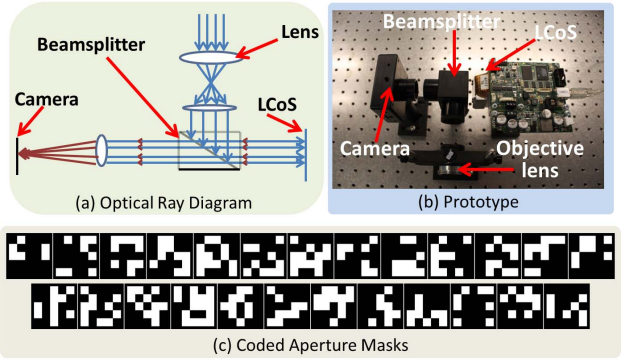

Figure 6. (a) Optical ray diagram of our setup. (b) Actual working prototype. (c) All 25 coded aperture masks used in our approach.

(since the maximum aperture size that our system can support is 10×10 mm) and group it into 25 2×2 mm sized pinholes to obtain 5×5 angular resolution. The LCoS equivalently acts like a mask with a pattern of zeros and ones, transmitting light where the LCoS pattern is one and blocking light where it is zero. This enables us to capture multiplexed angular views of the light field at a very high rate by simply changing the multiplexing pattern at the LCoS. The frames are captured at 25Hz using a pointgrey DragonFly2 camera.

6. Results

In this section, we show real results on challenging datasets captured using our prototype. Note that we capture a 25 fps video, where each captured frame is a coded aperture multiplexed image of the scene and our angular LF resolution is 5×5.

Digital Refocusing on Dynamic Scene: Figure 7 shows a real dataset consisting of several markers in front of a textured background. The orange colored EXPO marker is moving in the scene from right to left[2]. Figure 7 shows three captured multiplexed images at different time frames (frames 13, 88 and 330), along with the refocused images (front and back) for the entire scene. The object motion is ~ 0.22 pixels per frame, leading to overall motion of 70 pixels between frame 13 and 330. Notice that there are no artifacts in digital refocusing on the moving object. More importantly, refocusing results are obtained at *full* sensor resolution.

Reconstruction of Dynamic LF Views: However, the true merit of a LF video camera is in obtaining artifact free angular information for dynamic scenes. The angular LF information is referred to as LF 'views'. Each view is a sub-aperture image of LF. Note that for each time instant t, our approach reconstructs $5 \times 5 = 25$ LF views. The digital refocusing results combine information from all the LF views and thus artifacts in LF views could get suppressed in refocused images. Thus, LF views show the actual quality of reconstructed LF.

Figure 8 shows another real dataset consisting of a cup (black) moving from right to left and two static objects. In Figure 8, we show the captured image for frames 162 and 242 along with reconstructed top-left LF view $(L(1,1,s,t)$ at those time instant. We compare our motion-aware reconstruction with another reconstruction using fixed WL of $F = 25$ frames for *each* pixel. Since our approach decides the number of frames F to be used for each patch adaptively, it produces significantly better LF views. Notice that this comparison uses the exact same dictionary and sparse representation for both results. Thus, motion-aware adaption enables handling dynamic objects for LF reconstruction.

Figure 9 shows another scene with a moving toy in front of complex textured background. The input image and computed optical flow for frame 153 are shown along with digital refocusing (front and back) using reconstructed LF. Since the method of [18] completely ignores scene motion, it shows severe artifacts on the moving toy. Further, even in static regions, our sparse regularization results in better noise handling and consequently sharper reconstruction. The fixed window length reconstruction utilizes the same learned dictionary and sparse reconstruction, but results in low-resolution refocusing. Our motion-aware adaption enables high-quality refocusing.

7. Discussions and Conclusions

We presented a novel programmable aperture light field camera that exploits highly optimized coded aperture patterns and a dictionary learning/sparse representations based framework for high resolution LF reconstruction.

Most LF cameras suffer from a resolution trade-off resulting in significant loss of spatial resolution. Our method allows reconstruction of light-fields at the spatial resolution of the image sensor. Compared to previous programmable aperture based LF methods, we achieve a much higher temporal resolution on account of the motion-aware sparse regularized reconstruction algorithm. However, since our codes are 50% and because transparent polarization based LCOS modulators suffer from an additional 50% loss that DMD implementations do not suffer from, we end up with more than 75% light loss. Another demerit of our algorithm is that the reconstruction is slow. (10min/frame on a intel i7 2600 CPU for $600 \times 800 \times 5 \times 5$ LF). Finally, the motion aware reconstruction algorithm can only account for motion up to 1 px/frame, larger motions result in reconstruction artifacts, due to inadequately modelled motion blur.

Acknowledgements

S.T and A.V were supported by NSF Grants NSF-IIS:1116718, NSF-CCF:1117939 and by funding from a Samsung GRO research award.

[2]Supplementary materials contain captured video for all data sets.

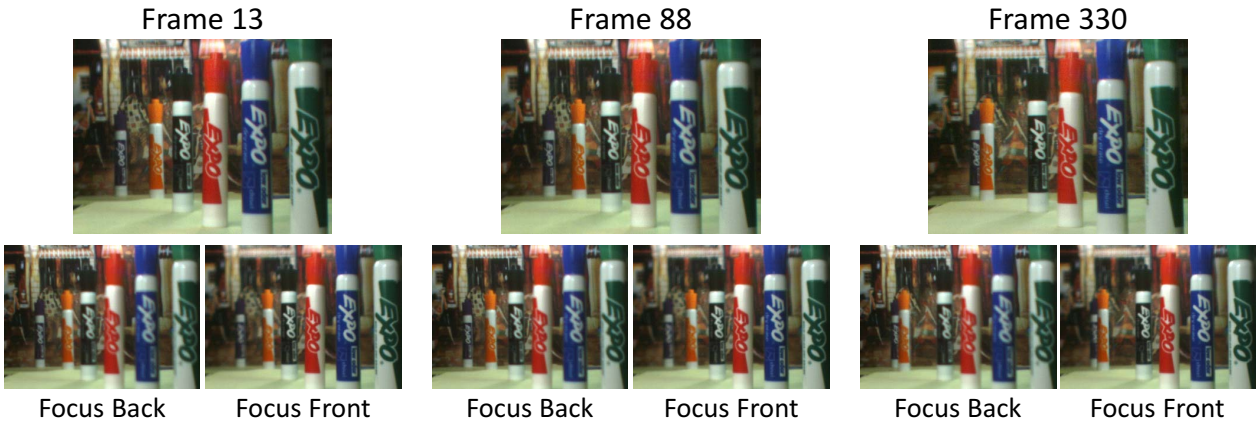

Figure 7. Motion-aware digital refocusing on moving objects. (Top) Three frames of the captured video using our setup shows the orange EXPO marker moving from right to left. The marker is moving *continuously*. (Bottom) Digital refocused images (front and back) corresponding to each time-frame. Notice that there are no artifacts on the moving object.

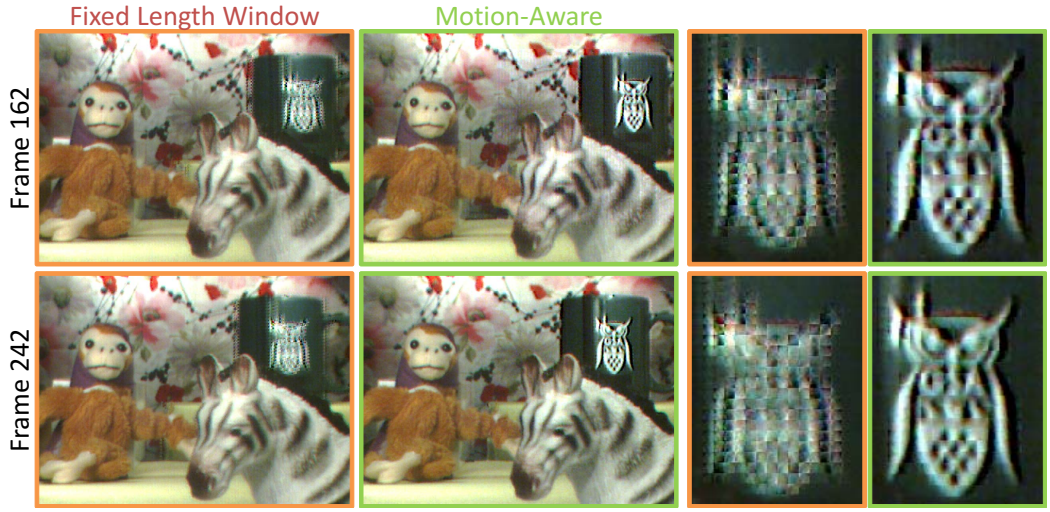

Figure 8. Comparison of our motion-aware reconstruction (using 25 frames) with reconstruction using fixed length window for each pixel ($W = 25$). Here we show the reconstructed bottom left LF 'view' for two frames. A zoomed in view of moving object clearly demonstrates that our motion-aware reconstruction successfully removes artifacts on the moving object.

References

[1] A. Agrawal, A. Veeraraghavan, and R. Raskar. Reinterpretable imager: Towards variable post-capture space, angle and time resolution in photography. *Computer Graphics Forum*, 29:763–773, May 2010. 3

[2] M. Aharon, M. Elad, and A. Bruckstein. K-svd: Design of dictionaries for sparse representation. *Proc. SPARS*, 5:9–12, 2005. 4

[3] D. Babacan, R. Ansorge, M. Luessi, P. Ruiz, R. Molina, and A. Katsaggelos. Compressive light field sensing. *IEEE Trans. Image Processing*, 21:4746–4757, 2012. 1, 2, 3

[4] T. E. Bishop, S. Zanetti, and P. Favaro. Light field superresolution. In *Proc. Int'l Conf. Computational Photography*, pages 1–9, 2009. 1, 3

[5] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008. 5

[6] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006. 3

[7] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Trans. Image Processing*, 20(7):1838–1857, 2011. 4

[8] D. L. Donoho. Compressed sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, 2006. 3

[9] J. M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Trans. Image Processing*, 18(7):1395–1408, 2009. 5

[10] M. Elad. Optimized projections for compressed sensing. *IEEE Trans. Signal Processing*, 55(12):5695–5702, 2007. 5

[11] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE*

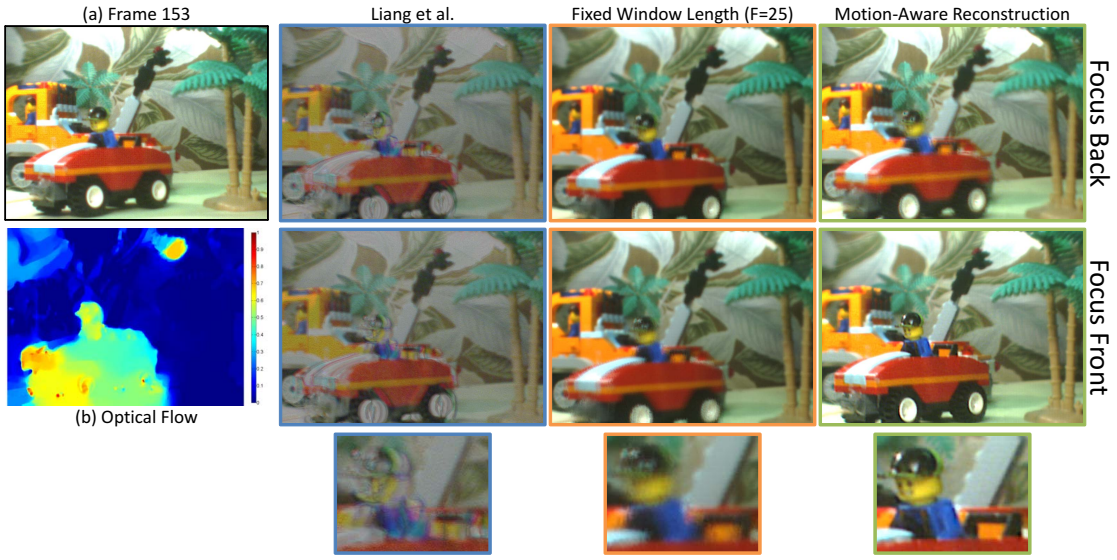

Figure 9. Comparison of Liang *et al*. [18] and fixed window length reconstruction with our motion-aware reconstruction. Notice that Liang *et al*. [18] completely fails on moving objects while our motion-aware approach enables high-quality refocusing.

Trans. Image Processing, 15(12):3736–3745, 2006. 4

[12] T. Georgiev and A. Lumsdaine. Superresolution with plenoptic camera 2.0. *Adobe Systems Inc., Tech. Rep*, 2009. 1, 2, 3

[13] T. Georgiev, C. Zheng, S. Nayar, B. Curless, D. Salasin, and C. Intwala. Spatio-angular resolution trade-offs in integral photography. In *EGSR*, pages 263–272, 2006. 2

[14] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar. Video from a single coded exposure photograph using a learned over-complete dictionary. In *ICCV*, pages 287–294, Nov. 2011. 4

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

[16] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas. Synthetic aperture confocal imaging. *ACM Trans. Graph.*, 23(3):825–834, 2004. 2

[17] M. Levoy and P. Hanrahan. Light field rendering. In *SIGGRAPH*, pages 31–42, 1996. 2

[18] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. Chen. Programmable aperture photography: Multiplexed light field acquisition. *ACM Trans. Graphics*, 27(3):55:1–55:10, 2008. 1, 2, 3, 4, 6, 8

[19] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, MIT, 2009. 5

[20] Lytro. The lytro camera. https://www.lytro.com/. 1, 2

[21] R. F. Marcia, Z. T. Harmany, and R. M. Willett. Compressive coded aperture imaging. In *Proc. SPIE*, 2009. 3

[22] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive Light Field Photography using Overcomplete Dictionaries and Optimized Projections. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 32(4):1–11, 2013. 2, 3

[23] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S. K. Nayar. Programmable aperture camera using lcos. In *ECCV*, pages 337–350, 2010. 2, 5

[24] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with handheld plenoptic camera. Technical report, Stanford U, 2005. 2, 3

[25] P. Peers, D. K. Mahajan, B. Lamond, A. Ghosh, W. Matusik, R. Ramamoorthi, and P. Debevec. Compressive light transport sensing. *ACM Trans. Graphics*, 28(1):3, 2009. 3

[26] Pointgrey. Profusion 25 camera. 1

[27] Raytrix. 3d light field camera technology. http://www.raytrix.de/. 1, 2, 3

[28] G. K. Skinner. X-Ray Imaging with Coded Masks. *Scientific American*, 259:84, Aug. 1988. 3

[29] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Information Theory*, 53(12):4655–4666, 2007. 4

[30] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69:1–69:12, 2007. 2, 3

[31] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, 2005. 2

[32] X. Yuan, J. Yang, P. Llull, X. Liao, G. Sapiro, D. J. Brady, and L. Carin. Adaptive temporal compressive sensing for video. *arXiv preprint arXiv:1302.3446*, 2013. 3

[33] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, pages 2691–2698, 2010. 4