Dictionary-based light field acquisition using sparse camera array

Xuan Cao,* Zheng Geng,* and Tuotuo Li

State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

*zheng.geng@ia.ac.cn

Abstract: We propose a dictionary-based dense light field acquisition technique. This technique captures light field successfully from a sparse camera array with no mask or any other optical modifications on cameras. Light rays in wider field are captured by our system to achieve larger disparity and higher angular resolution. We also accelerate the reconstruction of light field significantly by a local sliding window which applies median filter only in *disaster areas* and acquire satisfactory quality. In our experiments, light field with 7x7 views at resolution of 384x512 is restored from 5 cameras with PSNR of 33.0192dB with a computing time of 1.85 hours on a consumer-grade desktop computer.

OCIS codes: (110.0110) Imaging systems; (100.3020) Image reconstruction-restoration; (110.5200) Photography; (110.1758) Computational imaging.

References and links

1. J. Geng, "Three-dimensional display technologies," Adv. Opt. Photonics **5**(4), 456–535 (2013).
2. B. Wilburn, N. Joshi, V. Vaish, E. 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).
3. R. Ng, "Fourier slice photography," ACM Trans. Graph. **24**(3), 735–744 (2005).
4. R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Tech. rep., Stanford University, (2005). *http://graphics.stanford.edu/papers/lfcamera/lfcamera-150dpi.pdf*
5. M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," ACM Trans. Graph. **25**(3), 924–934 (2006).
6. www.raytrix.de
7. K. Venkataraman, D. Lelescu, J. Duparre, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "PiCam: An ultra-thin high performance monolithic camera array," ACM Trans. Graph. **32**(6), 166 (2013).
8. E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory **52**(2), 489–509 (2006).
9. J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," IEEE Trans. Inf. Theory **53**(12), 4655–4666 (2007).
10. M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. Image Process. **15**(12), 3736–3745 (2006).
11. M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *in Proceedings of the IEEE* (Special Issue on Application of Sparse Representation and Compressive Sensing, 2010), pp. 972–982.
12. A. Y. Yang, Z. H. Zhou, A. G. Balasubramanian, S. S. Sastry, and Y. Ma, "Fast l(1)-minimization algorithms for robust face recognition," IEEE Trans. Image Process. **22**(8), 3234–3246 (2013).
13. S. D. Babacan, R. Ansorge, M. Luessi, P. R. Matarán, R. Molina, and A. K. Katsaggelos, "Compressive light field sensing," IEEE Trans. Image Process. **21**(12), 4746–4757 (2012).
14. K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," ACM Trans. Graph. **32**(4), 46 (2013).
15. Q. Yao, K. Takahashi, and T. Fujii, "Compressed sensing of ray space for free viewpoint image (FVI) generation," ITE Trans. Media Tech. and App. **2**(1), 23–32 (2014).
16. M. DaneshPanah, B. Javidi, and E. A. Watson, "Three dimensional imaging with randomly distributed sensors," Opt. Express **16**(9), 6368–6377 (2008).
17. Y. Rivenson, A. Stern, and J. Rosen, "Compressive multiple view projection incoherent holography," Opt. Express **19**(7), 6109–6118 (2011).
18. E. J. Candes and T. Tao, "Decoding by linear programming," IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005).

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

19. A. Cohen, W. Dahmen, and R. Devore, "Compressed sensing and best k-term approximation," J. Am. Math. Soc. **22**(1), 211–231 (2009).
20. M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of ACM SIGGRAPH*, 31–42. (1996).
21. http://web.media.mit.edu/~gordonw/SyntheticLightFields/
22. J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "On-line dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, (Montreal, Canada, 2009), pp. 689–696.
23. J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," J. Mach. Learn. Res. **11**, 19–60 (2010).

1. Introduction

In recent years, light field acquisition has gained considerable attraction to facilitate wider-spread applications of three-dimensional (3D) imaging and display technologies [1]. Two problems still hinder the development of light field acquisition technologies: 1) A large number of cameras are needed to capture light field. 2) If the number of cameras used in an array is limited, the disparity among different views becomes small and angle resolution becomes low. Generally, light field acquisition with higher angle resolution and larger disparity needs more cameras. One has to make a well-balanced trade-off between the number of cameras and disparity of light field. Specifically, a fully populated camera array is often needed to capture all views from different perspectives [2]. Even with the decreasing unit cost of cameras, an array of cameras with a moderate size (e.g. 7x7) would require significant number of cameras (e.g. 49), resulting in a complex system configuration and expensive hardware cost set up for a practical light field acquisition system. Commercial light field cameras are now available, such as lytro [3–5], raytrix [6], and pelican [7], which significantly reduce the number of cameras and tend to have small package size. However, they have very limited separation among multiple sensors, resulting in very limited light field disparity. This, among other things, has prevented their uses to a wider variety of commercial application.

How to break the above-mentioned trade-off, and to capture a high angle resolution light field with large disparity from very limited number of cameras? Sparse representation [8,9] provides a novel solution to this problem. As a new approach for information restoration from limited measurements, sparse representation has been applied successfully in many fields, such as image processing [10,11], face recognition [12] and light field acquisition [13,14]. Due to the fact that most signals in our practical application are not sparse, generally a redundant basis set is constructed to represent the original signals sparsely. Being superior than commonly used bases, a novel set of basis learned from the training data can have more meaningful and sparser representation ability [10,14,15]. The whole set of learned basis is stored in an over-complete dictionary. Each column vector in the dictionary is a basis which is called an atom. Both Marwah [14] and Yao [15] demonstrated that 3D scene information can be restored from a dictionary with higher quality.

In this paper, a sparse camera array (SCA) is constructed to acquire full light field using sparse representation and redundant dictionary. We assume that light rays in different directions emitted from an identical space point are different, details in section 2.1. We found the light rays' distribution in our model follows certain patterns-light field atoms, which can represent the light field patch by limited number of bases combination. This sparsity assures that light field can be restored successfully with a high probability.

Experimentally, two issues needed to be handled. 1) The richer patterns in dictionary, the higher probability to predict the unknown light field patch. 2) Stronger correlation between the measurement from camera array and the patterns from light field dictionary brings higher ratio to choose the 'right' bases. The first issue depends on the diversity of training data, we trained a dictionary associated with all color channels because light fields for training have different patterns distributed at all color channels, see section 3.1. The second issue involves the solving of l_0 norm minimization problem. The sparsity of representation could not be hold for all light field patches. Some patches are reconstructed with poor quality even though they are represented by combination of numerous atoms. We call the degraded patch reconstruction

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

disaster area, details in section 3.3. We detect and improve the *disaster areas* by exploring the relationship among PSNR (Peak Signal to Noise Ratio) of the reconstruction, PSNR of the measurement and sparse level, see section 3.3.

1.1 Related work

Light field reconstruction from sub-sampling emerged in recent years. Babacan placed a randomly coded mask at the aperture of a camera and restored the light field images by Bayesian reconstruction algorithm [13]. Marwah adopted an optional mask which was mounted slightly on the camera sensor and reconstructed the light field from a light field dictionary [14]. These two designs mixed and compress the target signals (coding) so multi-views can be captured (decoded) from a single camera. Despite various advantages of these designs, disparity of the acquired light field, however, is limited by the size of optical comments. Adding a mask or other optical designs could also be a burden. Additionally, the mask reduced the light transmission which resulted in lower SNR (Signal to Noise Ratio). DaneshPanah adopted randomly distributed sensors to fulfill 3D imaging [16]. Rivenson reconstructed 3D scene from the randomly sub-sampled generated Fourier hologram [17]. Both DaneshPanah and Rivenson demonstrated that 3D scene information can be restored from spatially randomly sub-sampling. Leveraging these previous contribution, and in contrast to their technical approach, we propose a novel light field acquisition approach using a sparse camera array. We combine the light field dictionary and spatially randomly sub-sampling architecture, and our approach eliminates the need for a mask and allows acquisition of light field with larger disparity.

1.2 Main contributions

We reconstruct the light field from a limited number of unmodified cameras and accelerate the algorithm for light field reconstruction. Particularly, we make the following contributions:

a) We propose a sparse camera array architecture to capture the full light field and train an over-complete dictionary containing rich light distribution patterns (atoms) to represent light field by combining limited number of atoms.

b) Sparse camera array simplifies the light field camera design, eliminating the need for a mask and acquiring light field with much larger disparity. As an example, we demonstrate that the fully populated 7x7 light field could be recovered successfully from only 5 cameras by exploring reusability of light field patterns, even though measurement matrix doesn't seem to hold the RIP (Restricted Isometry Property [18,19]) well.

c) We detect *disaster areas* by analyzing the relationship among PSNR of reconstruction, PSNR of measurement and sparse level. We then accelerate the light field reconstruction by applying local sliding window only on the *disaster areas*, details in section 3.3.

2. Model of light field acquisition

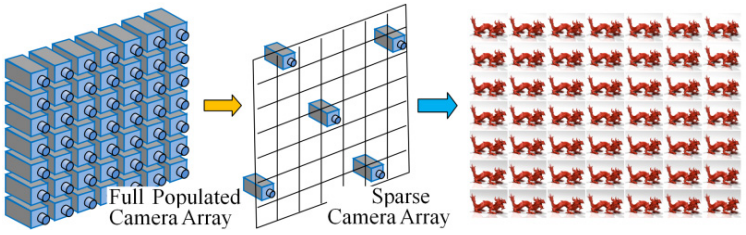


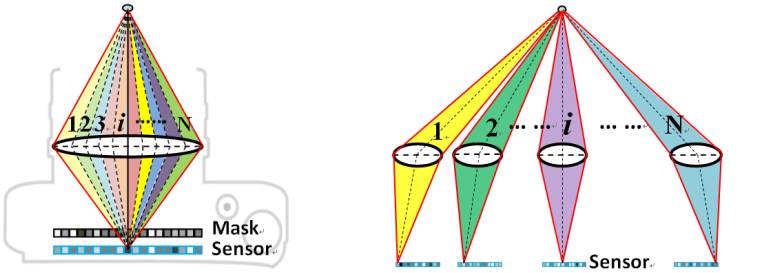
Fig. 1. Schematic diagram of our architecture.

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

In our works, we follow the definition of light field as a 4D function [20] and the schematic diagram is shown in Fig. 1. The 7x7 camera array is reduced to 5 cameras which capture the full 49 light field views successfully.

2.1 Eliminate the need for mask and achieve larger disparity

Some previous light field acquisition technologies depend heavily on using a coded mask [13,14], as shown in Fig. 2(a). For any point in objective space, one pixel on camera sensor records the weighted sum of all light rays passing the aperture and modulated (weights) by the mask. The target light field is restored by decoding the mixed signals recorded on camera sensor. Obviously, only light rays passing through the aperture can be reconstructed. Therefore, the upper bounder of disparity is restricted by the size of aperture and sensor. Furthermore, a coded mask with both high transmittance and good separation is difficult to build, and the modification on optics and calibration on camera increase the burden.



(a) Light field acquisition with mask (b) Light field acquisition by camera array

Fig. 2. Comparison between single camera with mask and camera array.

We model the light field acquisition as Fig. 2(b) shows. We assume that light rays emitted from one space point in different direction are different. Our following experiments demonstrate that this assumption is reasonable. As shown in Fig. 2(b), the light rays are recorded discretely by N*N cameras in both horizontal and vertical dimensions. As there is no modification on each camera, we don't need to rack our wits about the mask design issues. Without the mask, higher light transmission brings higher SNR. The disparity of reconstructed light field is determined by the relative distance among cameras which is generally larger than the aperture size. In this paper, we will demonstrate a novel technique to restore the full N-by-N light field from a limited number of unmodified cameras based on an over-complete dictionary representing the light field sparsely.

2.2 Measurement matrix analysis

As shown in Fig. 2(b), the coded mask which plays a key role in previous measurement matrix has been eliminated in our model. We now analyze the projection of light field from full camera array to sparse camera array. N*N views in light field are vectorized (N = 7 in this case), and the M images captured by sparse camera array are vectorized in the same way (M = 5 in this case). The measurement matrix, as illustrated in Fig. 3, is very sparse: only the diagonal elements in M sub-matrixes are non-zero. If different views are captured, sub-matrixes are placed in different position. The measurement matrix maintains global sparsity and local aggregation: the limited non-zero items fail to evenly distribute in whole measurement matrix but to cluster in M sub-matrixes. It's very difficult to capture the global information of target light field by such a non-globally distributed measurement matrix.

RIP (Restricted Isometry Property) is very useful in studying the general robustness as to whether a target signal could be restored from limited measurements. If the measurement matrix Φ holds Eq. (1) for all k - sparse signal x when δ_k is not too close to one, Φ obeys the RIP of order k [18, 19].

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

$$(1-\delta_k)\,\|x\|_{\ell_2}^2 \le \|\Phi x\|_{\ell_2}^2 \le (1+\delta_k)\,\|x\|_{\ell_2}^2 \tag{1}$$

where k is the sparse level, and Φ, x are measurement matrix and sparse signal, respectively.

$$m \ge C \times k \times \log(\frac{n}{k}) \tag{2}$$

If RIP could be hold at order k, then m measurements are needed to recover the original sparse signal as shown in Eq. (2), where n is the total number of atoms in dictionary and C is positive constant. It seems to be very complicated to find a pairwise sparse-level (k) and number-of-measurements (m) to enforce the RIP condition.

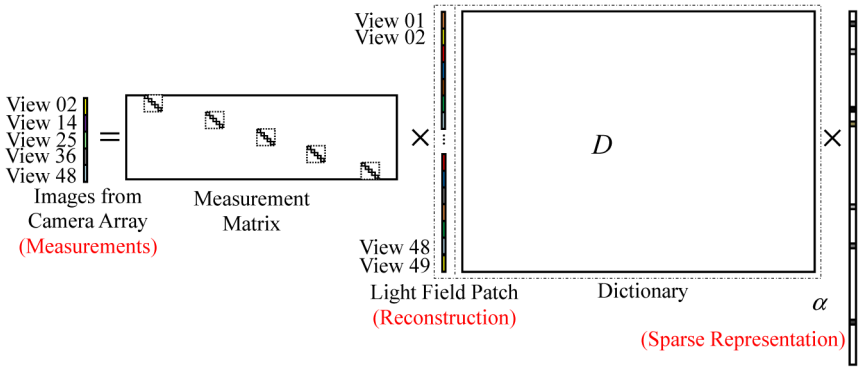


Fig. 3. Matrix representation of light field photograph from sparse camera array.

Regarding the local aggregation of the measurement matrix, RIP may not be hold well. It's therefore difficult to restore the full light field from limited measurements. This is the main challenge we are facing in this paper. Although the measurement matrix in our model does not seem to fully comply with the RIP condition, we found that the light field patterns have good repeatability. We demonstrated experimentally that the full light field could be restored quite well by exploring this repeatability. Section 3.1 and 3.2 offer more details.

2.3 Distribution of camera array

The spatial distribution of sparse camera array determines the positions of sub-matrixes in measurement matrix. In order to acquire the global information of light field as much as possible, it's preferable to place cameras in somewhat evenly distributed patterns. We test several distribution patterns with only 5 cameras, as shown in Fig. 4. The PSNR achieved by these four camera distribution patterns (from left to right) are 30.4031dB, 28.5886dB, 29.7695dB and 31.7844dB. We finally choose the forth distribution pattern for its wide span and with no duplicated cameras in each row or column.

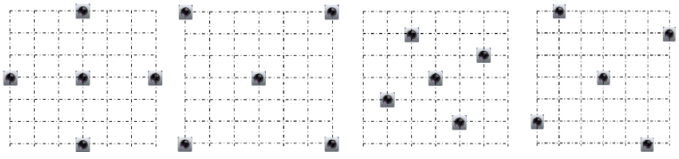


Fig. 4. Distribution of cameras.

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

3. Light field capture and acceleration

3.1 Dictionary training

The light rays' distribution of one pixel follows certain patterns, but this inter-pixel coherence seems defective to restore the full light field from limited measurements. The among-pixels coherence is also explored by extracting pixel block ($p_x \times p_y$) parallel in all views ($p_u \times p_v$). Thus the size of light field patch is $n = p_x \times p_y \times p_u \times p_v$. By exploring this among-pixels coherence, more 3D information, including parallax, occlusion and focus, can be learned and stored in light field dictionary. However the training data pool becomes very huge and partial training data are not meaningful. To solve this problem, we remove flat patterns, such as the pure background which has no difference from view to view, by simply figuring out the residual error between the first and last view.

Diversity of training data is crucial to our dictionary training since richer patterns of training data contain more useful light field behaviors information. We found that some patterns appeared in one color channel could be used to make up pattern deficiency in other color channels. From another perspective, patterns in light field reflect the spatial relationship among the spatial points in real world, such as parallax, obscuration information and depth cues. Obviously, these spatial relationships could exist in any color channel. So we mix light field patches in three color channels into the one training data pool.

After filtering flat patterns and mixing all color channels, a large set of patches is extracted as training data and a gray dictionary $\mathcal{D} \in \mathbb{R}^{n \times d}$ is learned as the Eq. (3).

$$\underset{\{\mathcal{D},\mathcal{A}\}}{minimize} \left\| \mathcal{L} - \mathcal{D}\mathcal{A} \right\|_F , \quad s.t. \ \forall j, \left\| \alpha_j \right\|_0 \le k \tag{3}$$

where $\mathcal{L} \in \mathbb{R}^{n \times q}$ is the training data including q light field patches and $\mathcal{A} = [\alpha_1, \alpha_2, \ldots\ldots, \alpha_q] \in \mathbb{R}^{d \times q}$ is the sparse coefficients matrix, each column in $\mathcal{A}$ is a sparse vector which contains at most k non-zero items ($k \ll d$). There is no non-negative restriction in our dictionary training allowing atoms in our light field dictionary to be negative.

Proper visualization of trained dictionary may offer us intuition on the diversity of data set. As shown in Fig. 5, each atom seems meaningful, containing different parallax, obscuration and focus information. In other words, the patterns in our dictionary are very rich.

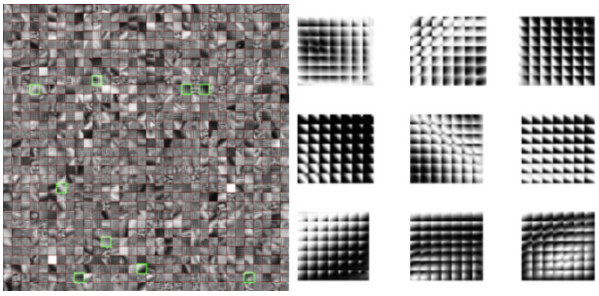


Fig. 5. Visualization of trained dictionary.

3.2 Light field reconstruction

Light field reconstruction is the inverse problem of dictionary training. As shown in Eq. (4), firstly we figure out the sparse coefficient α for each light field patch. Then we reconstruct the patch by combining atoms in the trained dictionary according to the representation vector.

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

$$\begin{cases} \min\limits_{\alpha} \ \left\| m - \Phi \mathcal{D} \alpha \right\|_2 + \lambda \left\| \alpha \right\|_0 \\ P_r = \mathcal{D} \alpha \end{cases} \tag{4}$$

where m is measurement of current patch, α is the sparse coefficient, Φ is the measurement matrix determined by the spatial distribution of sparse camera array, $\mathcal{D}$ is the trained dictionary and P_r is the reconstruction of current patch.

λ is sparsity penalizing parameter determining the sparse level of α. Larger λ brings sparser level of α which means combination of smaller number of atoms. Our experiments showed that, for most patches, either too many or too few atoms' combination will degrade the quality of reconstruction. Thus an appropriate λ will help most patches to achieve satisfactory reconstruction quality.

It's difficult to find an appropriate λ for both different dictionaries and different testing data. After finishing dictionary training, we tested several values of λ in our experiments. Table 1 shows the reconstruction quality for sparse camera array with 5 or 9 cameras to restore 7x7 light field.

Table 1. Reconstruction PSNR(dB) under different λ

	λ = 0.01	λ = 0.006	λ = 0.005	λ = 0.004	λ = 0.003	λ = 0.002
5 cameras	31.5946	31.7567	31.7811	31.7844	31.7573	31.6624
9 cameras	36.3650	36.5534	36.5701	36.5601	36.5001	36.3315

3.3 Acceleration: local sliding window only in disaster areas

Dictionary training is sensitive to the given parameters. We can train the dictionary based on the sparse level as shown in Eq. (3) and we can also define the priority to error tolerance on the representation. The PSNR of some reconstructed patches maybe very low if the sparse level was given more priority when training dictionary, because combination of too fewer atoms may not sufficient to represent the complicated patches well. Therefore it seems extremely difficult to train a dictionary which performs very well on reconstructing every light field patch, no matter is the sparse level or representation error tolerance given priority in dictionary training. We define the degraded patch as disaster area of reconstruction, see the mosaic pixel blocks shown in Fig. 6(a), and the enlarged views are shown in Fig. 6(b). Specifically, the disaster area is the reconstructed pixel block with relatively lower PSNR. Disaster *areas* contribute significantly to the global degraded reconstruction quality. Furthermore, the *disaster areas* have very low regional PSNR of reconstruction which seriously degrade visual comfort level.

In the stage of light field restoration, traditional strategy is to construct a distinct window restoring the light field by discrete pixel blocks without overlapping pixels. As Fig. 6(b) shows, distinct window fails to capture the parallax and occlusion information and results in obvious *disaster areas*.

Global sliding window, as another strategy, merges the overlapping patches with a median filter or mean filter [14]. The size of sliding window is same with that of the pixel block ($p_x \times p_y$) in extracting training data. The global sliding window scans the entire light field in a fixed step. A smaller step results in more overlapping pixels, and vice versa. More information, especially parallax and occlusion, could be restored from dictionary by a sliding window. However computational cost increases multi-folds under the strategy of global sliding window. In the case of a 7x7x384x512 light field with three color channels, totally (384-8)x(512-8)x3 = 568,512 patches need to be reconstructed (patch size is 7x7x8x8), while only (384/8)x(512/8)x3 = 9216 patches need to be recovered by distinct window. Approximately, the computing time using strategy of sliding window is (568512/9216) = 61.7 times longer than that under distinct window. Reconstructing a light field of 5x5x480x270 mentioned in [14] consumed 18 hours with median filter using the global sliding window.

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

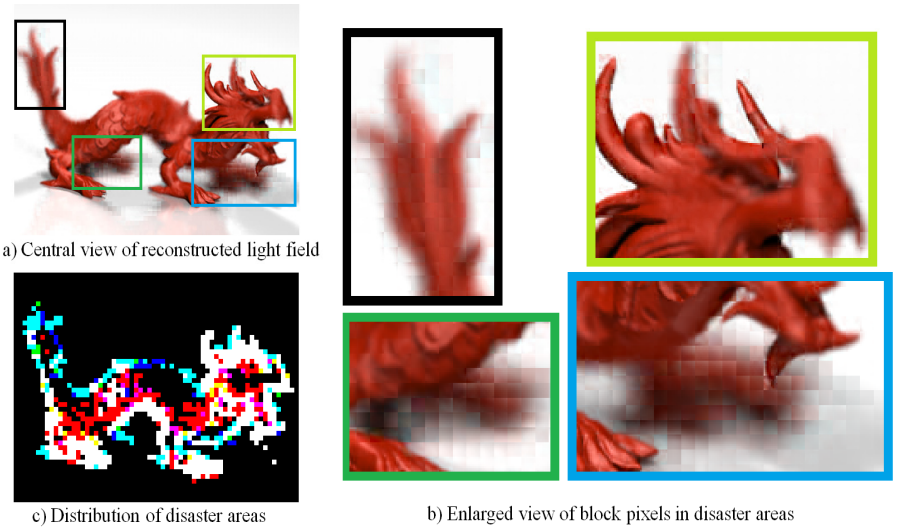


Fig. 6. Reconstruction of light field under distinct window and visualization of *disaster areas*.

According to the statistic data shown in Fig. 7, most reconstructed patches achieve satisfactory PSNR under the strategy of distinct window. Only a small number of patches meet criteria of *disaster area*. So we can apply a local sliding window only in *disaster areas*, improving the degraded patches with median filter and saving significant time.

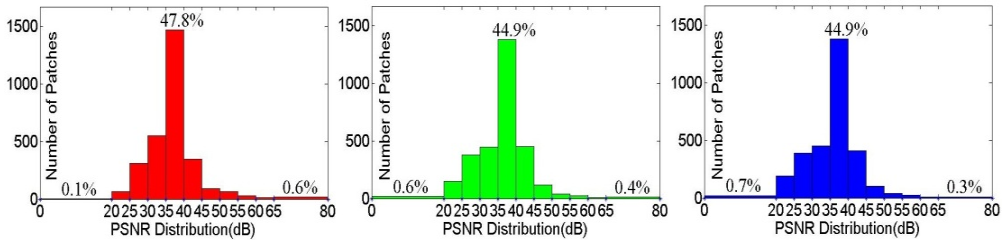


Fig. 7. Statistic data of reconstruction quality under distinct pixels block. *The red, green and blue histograms represent the statistic PSNR of reconstructed patches in RGB color channels, respectively.

How can we detect the position of *disaster areas*? PSNR of reconstruction is a good criteria to detect the degraded pixel blocks, but the PSNR could not be calculated until the full target light field is known. We found that the PSNR of measurement and sparse level of representation can reflect the distribution of PSNR of reconstruction well. Figure 8 shows the distribution of three criteria (see red labels shown in Fig. 3): 1) PSNR of reconstruction- PSNR_{P_r}, 2) PSNR of measurement- PSNR_{m_r}, 3) Sparse level of representation- S_{α}. These criteria are calculated out as Eq. (5) - Eq. (7):

$$\begin{cases} \begin{cases} \min\limits_{\alpha_{(i)}} \left\{ \left\| m_{(i)} - \Phi \mathcal{D} \alpha_{(i)} \right\|_2 + \lambda \left\| \alpha_{(i)} \right\|_0 \right\} \\ P_{r(i)} = \mathcal{D} \alpha_{(i)} \end{cases} \\ \mathrm{PSNR}_{P_{r(i)}} = 10 \log_{10} \left(\dfrac{n}{\sum_{j=1}^{n} \left(P_{r(i)j} - P_{(i)j} \right)^2} \right) \end{cases} \tag{5}$$

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

where $P_{r(i)}$ is the reconstruction of *i*-th patch, $P_{(i)}$ is *i*-th target light field patch, n is the length of patch vector, j is the index of *j*-th term in the patch vector.

$$\begin{cases} \begin{cases} \min\limits_{\alpha_{(i)}} \left\{ \left\| m_{(i)} - \Phi\mathcal{D}\alpha_{(i)} \right\|_2 + \lambda \left\| \alpha_{(i)} \right\|_0 \right\} \\ m_{r(i)} = \Phi\mathcal{D}\alpha_{(i)} \end{cases} \\ \mathrm{PSNR}_{m_{r(i)}} = 10\log_{10}\left(\dfrac{n}{\sum_{j=1}^{l}\left(m_{r(i)j} - m_{(i)j} \right)^2} \right) \end{cases} \tag{6}$$

where $m_{r(i)}$ is the reconstruction of *i*-th measurement, $m_{(i)}$ is given *i*-th measurement; l is the length of measurement vector ($l < n$). It should be noted that the image pixels are stored as floating point data, so the peak value of pixels is 1.

$$\mathrm{S}_{\alpha_{(i)}} = \left\| \alpha_{(i)} \right\|_0 \tag{7}$$

where $\alpha_{(i)}$ is the sparse representation of *i*-th patch and $\mathrm{S}_{\alpha_{(i)}}$ is the number of non-zero terms in $\alpha_{(i)}$.

Experimentally, the distribution of PSNR_{P_r} follows two rules: 1) PSNR of reconstruction is not necessarily high if the PSNR of measurement was high. But a very low PSNR of measurement generally results in very low PSNR of reconstruction. 2) Generally, a sparse representation containing very small number of non-zero terms corresponds to high PSNR of reconstruction and vice versa. Firstly, it's almost impossible to achieve high reconstruction quality from degraded measurement. Low PSNR of measurement reflects that the sparse representation failed to represent the measurement well. So it's reasonable to doubt that the current sparse representation may also fail to represent the current patch reconstruction. Secondly, a good representation of measurement might not necessarily represent the target patch well. From the perspective of dictionary training, this situation may happen because the disappointing reusability of atoms in dictionary could not correctly 'guesses' the target patch. Finally, combination with fewer atoms implies that the target patch has been represented under very limited error. If not, more atoms combination is needed to decrease the error of reconstruction.

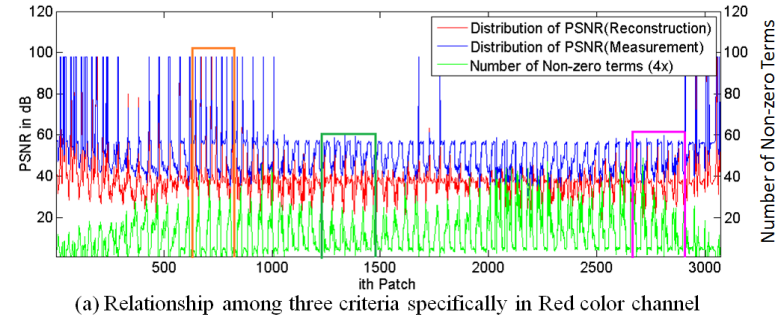
As shown in Fig. 8(b), (c) and (d), if one patch has a relatively lower PSNR of measurement and its representation has relatively larger number of non-zero term, this patch will have poor reconstruction quality. In this paper, *disaster areas* are detected separately in three color channels by setting thresholds on PSNR of measurement and sparse level of representation as Eq. (8). If the patch meets both thresholds' restriction at the same time, this patch will be regarded as *disaster area*.

$$\begin{cases} \mathrm{I_m} = \left\{ i \mid \mathrm{PSNR}_{m_r(i)} < \mu_m \times \dfrac{1}{T}\sum_{i=1}^{T} \mathrm{PSNR}_{m_r(i)} \right\} \\ \mathrm{I_\alpha} = \left\{ i \mid \mathrm{S}_{\alpha(i)} > \mu_\alpha \times \dfrac{1}{T}\sum_{i=1}^{T} \mathrm{S}_{\alpha(i)} \right\} \\ \mathrm{I_{disaster}} = \mathrm{I_m} \bigcap \mathrm{I_\alpha} \end{cases} \tag{8}$$

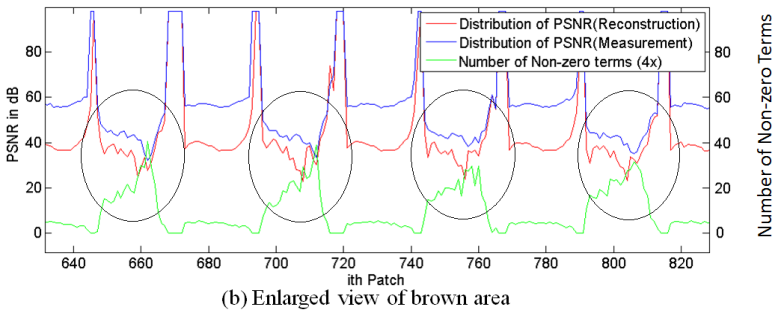
where $\mathrm{PSNR}_{m_r(i)}$ is PSNR of the *i*-th patch of measurement and $\mathrm{S}_{\alpha(i)}$ is the number of non-zero terms of *i*-th sparse representation vector. μ_m and μ_α are the thresholds factor

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

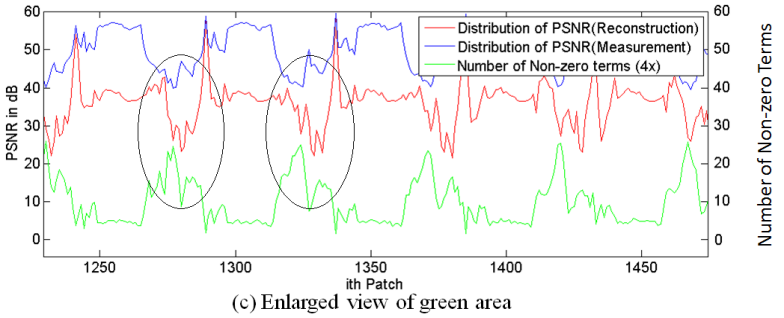
which function on $\mathrm{PSNR}_{m_r(i)}$ and $\mathrm{S}_{\alpha(i)}$. T is the total number of patches in the light field. The index of *disaster areas*— $\mathrm{I}_{\mathrm{disaster}}$ is the intersection of I_{m} and I_{α}.



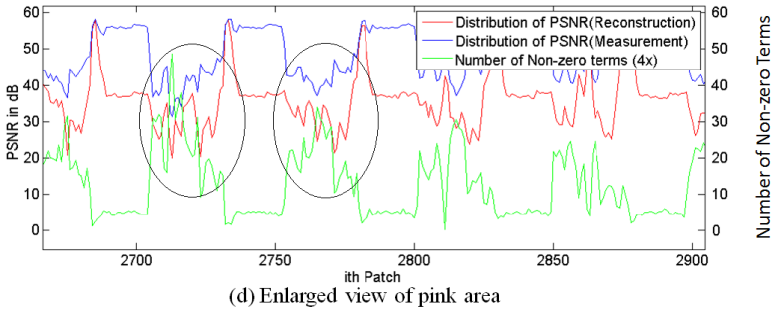
(a) Relationship among three criteria specifically in Red color channel



(b) Enlarged view of brown area



(c) Enlarged view of green area



(d) Enlarged view of pink area

Fig. 8. Relationship among PSNR of reconstruction, PSNR of measurement and sparse level
*The vertical axis on left side is for the PSNR of reconstruction and measurement in dB. The vertical axis on right side is for the number of non-zero terms. The actual number of non-zero terms is four times than that plotted in Fig. 8.

As shown in Fig. 6(c), most *disaster areas* are detected correctly. Red, green and blue blocks represent the *disaster areas* in red, green and blue color channels, respectively. And the

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

white blocks mean that corresponding areas meet reconstruction disaster in all three color channels. Other color blocks, such as cyan, yellow and crimson, represent the *disaster areas* in two color channels under the principle of color mixing. Totally 660, 579 and 612 disaster blocks are detected in red, green and blue color channels, correspondingly, when $\mu_m = 0.85$ and $\mu_\alpha = 1.4$.

More *disaster areas* will be detected if we relax the thresholds, and vice versa. We apply local sliding window around each *disaster area* and the effective diameter is two times larger than patch size. If the step of the local sliding window is 2 pixels, then $(660+579+612)\text{x}(\frac{16-8}{2})\text{x}(\frac{16-8}{2}) = 29616$ patches need to be reconstructed which costs much less computation resource than reconstruction of 568,512 patches under the strategy of global sliding window.

Table 2. Disaster improving under different steps and thresholds

$[\mu_m, \mu_\alpha]$	Disaster patches(RGB)	Step=1pixel		Step=2 pixels		Step=4 pixels	
		PSNR (dB)	Time (s)	PSNR (dB)	Time (s)	PSNR (dB)	Time (s)
[1,1]	[1257,1238,1246]	33.1088	19650	32.9794	4794	32.5328	2906
[1.4,0.85]	[660,579,612]	32.8661	11086	32.7854	3350	32.4395	1515
[1.2,0.8]	[338,305,348]	32.5091	8068	32.4823	2372	31.7844	987

As shown in Table 2, shorter step and more relaxed thresholds' restriction produce more *disaster areas* costing more time in reconstruction but achieving higher reconstruction quality. We have to perform a trade-off between the time cost and reconstruction quality. We set step as 2 pixels, which also save computer's RAM significantly than with a 1 pixel's step.

4. Implementation

Without any mask, we are relieved from the tasks of camera hardware modification and complex optical calibration. Images from the sparse camera array are all required measurements to recover the light field with the learned dictionary.

4.1Dictionary learning

As discussed in the section 3.1, we don't train dictionary independently in each color channel. Light field patches in all color channels are put into one training data pool to make sure that patterns appeared in one color channel could be reused in another color channel. The original light field data come from Camera Culture Group, Media Lab in MIT [21] and are divided into training data set and testing data set randomly. Training data set contains five light fields at resolution of 7x7x384x512 (7x7 viewpoints, resolution of each view is 384x512), and the patch size is 8x8. We have also tried different patch sizes, such as 6x8, 7x7, 9x9. While the final reconstruction PSNR under different patch sizes are very close and the deviation is kept in 3%. Relatively, a little higher PSNR is achieved by 8x8 patch size. Totally 1,110,375 patches are extracted from three color channels. After a randomly sampling of 60% and removing flat patterns as discussed in section 3.1, the training data pool contains 605,520 patches of training data. The size of dictionary significantly affects the reconstruction quality and computation time. Too large redundancy fails to bring performance improvement but to result in greater computation cost. We tested several discrete redundancy factors, including 1.3, 1.6, 1.7, 1.8 and 2. Twenty percent of patches in the target 4D light field were reconstructed. The average reconstruction quality (PSNR) under redundancy factor of 1.7 is highest. We finally choose 1.7 as redundancy factor, so the dictionary has 5332 atoms and each atom has 3136 elements. We employ online dictionary learning [22,23] as the Eq. (9).

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

$$\underset{\{\mathcal{D},\mathcal{A}\}}{minimize}\left\|\mathcal{L}-\mathcal{D}\mathcal{A}\right\|_F, \quad s.t.\ \forall j, \left\|\alpha_j\right\|_0 \le 1 \tag{9}$$

As mentioned in section 2.2, the spatial distribution of cameras determines the measurement matrix which fails to hold RIP well. Satisfactory reconstruction quality could not be achieved from the limited measurements (5 cameras). We expect each atom in dictionary has more "powerful" representation capability. The light field patch can then be represented as fewer atoms as possible. In the best condition, light field can be represented by one of atoms in dictionary. We therefore force the sparse level to be 1 in dictionary training stage, as the Eq. (9) express. Intuitively, this training strategy seems to help each atom in dictionary to learn more "essential" features and patterns about light field.

Although we force the training data set to be represented by one atom in dictionary training stage. But in reconstruction stage, most light field patches will be represented by multiple atoms' combination. Our experiments showed that this training strategy restores most patches with satisfactory PSNR, see statistic histogram shown in Fig. 7. What sacrificed is that performance in minority patches with serious reconstruction "disaster". We applied other special processing in these *disaster areas*, see local sliding window in section 3.3. Dictionary is learned on a workstation computer equipped with 32-core 2.6GHz Intel Xeon(R) CPU and 32 GB RAM in12.48 hours.

4.2 Light filed reconstruction

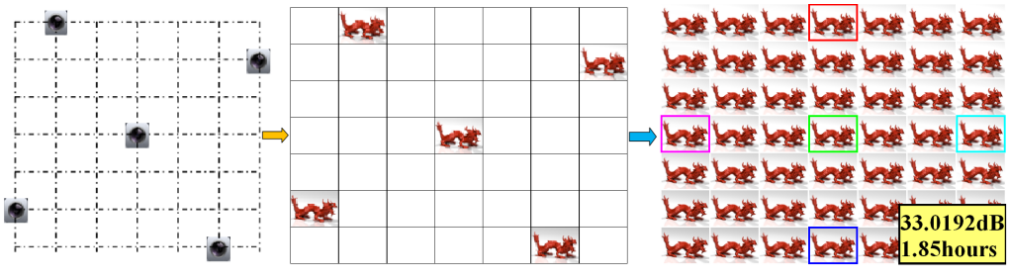
Larger number of cameras is available, easier to reconstruct light field. We test our system under 5 cameras' array and 9 cameras' array independently. We distribute 5 cameras as shown in Fig. 9(a). Distinct window is adopted as the main strategy to recover the light field patch by patch. The sliding window is only applied on *disaster areas*. The processing on disaster area contains two steps: 1) A local sliding window is applied around each disaster area (the sliding step is 2 pixels and the size of sliding window is 8-by-8 in this case). For one disaster area, $(\frac{16-8}{2})\times(\frac{16-8}{2})=16$ patches are extracted and each patch is reconstructed independently. Obviously, these patches have large overlapping areas. So pixels under the overlapping areas will get multiple pixel values as candidates. 2) By applying a median filter on every single pixel, the median value is chosen as the final pixel value.

We employ the SPArse Modeling Software [22,23] to solve the L_0 problem in Eq. (4) on a desktop computer equipped with 8-core 3.2GHz Intel(R) Xeon(R) CPU and 8G RAM. Sparsity penalizing parameter λ is set as 0.004. For 5 cameras, μ_m is 1.5 and μ_α is 0.3. Totally $3072\times3=9216$ patches are processed under distinct window. And 65050(R), 66300(G), 65350(B) patches are processed under sliding window. For 9 cameras, it's easy to achieve good reconstruction quality. So we enhance the thresholds' restriction as μ_m is 1 and μ_α is 1, resulting in 34575(R), 34375(G), 34725(B) disaster patches needing to be processed.

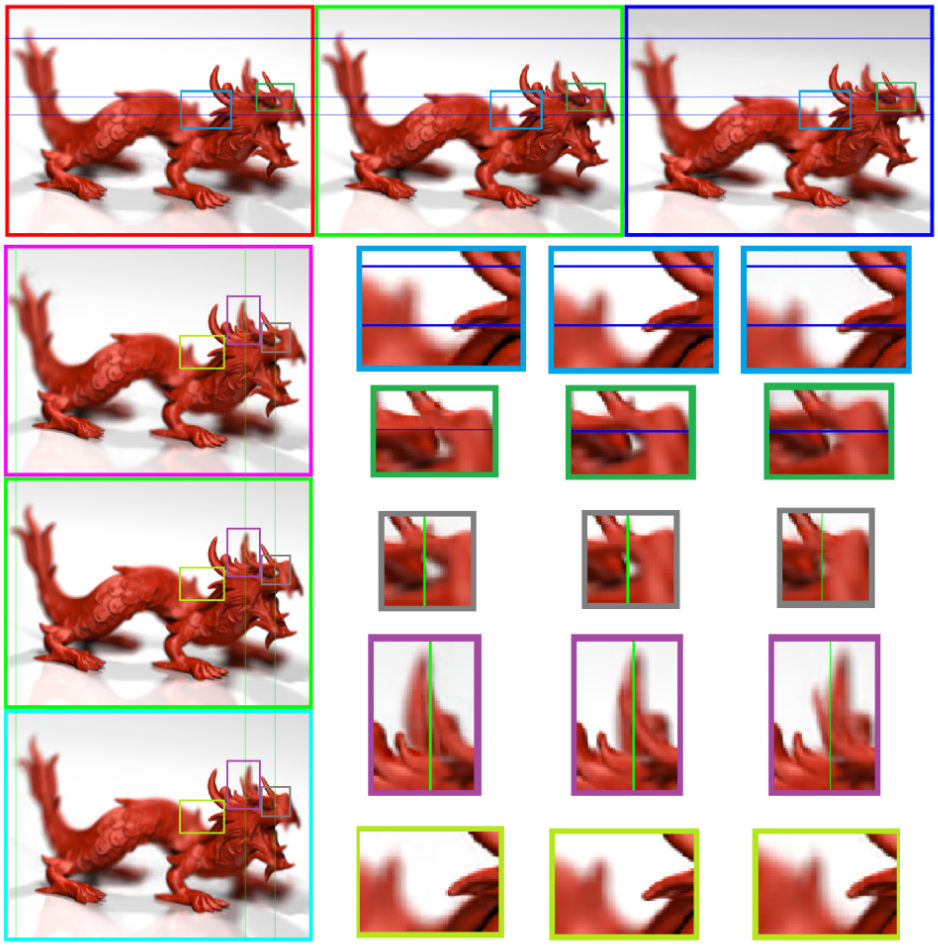
As shown the right side in Fig. 9(a), we found that different views in the light field have different reconstruction quality. Better quality is achieved when the view is closer to the given cameras. Conversely, views located far away from any given cameras achieve relatively poorer reconstruction quality. In the Fig. 9 (b), we enlarge the views far away from the given cameras to show the details of the reconstructed results. Other views closed to given cameras are not enlarged in this paper, but better reconstruction quality can be guaranteed.

Comparing the enlarged views in Fig. 6(b) and Fig. 9(b), the degraded mosaic pixel blocks are fixed by our local sliding window. As shown in Fig. 9(b), both the focused and out-of-focus areas are restored successfully, such as blue and yellow enlarged views. The horizontal and vertical parallax is large and clearly visible, as shown in blue, green and gray enlarged views. Additionally, occlusion information is also restored, see purple enlarged view. The PSNR of the full light field is 33.0192dB and the reconstruction is finished in 1.85 hours (the original light field data source is from [21]).

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014



(a) Architecture of light field photograph by 5 cameras



(b) Horizontal and vertical parallax and occlusion

Fig. 9. Distribution of 5 camera array and the reconstruction of light field.

Table 3 compares our design with similar architecture in Marwah's work [14]. The architecture in [14] is a single camera with mask, while ours is a sparse camera array without any modification on cameras. Although multi-cameras are needed in our design, there is no extra optical design and modification for cameras. More importantly, we are able to achieve much larger disparity. Our architecture can be easily constructed with off-the-shelf cameras. With the cost of camera unit decreasing, our design offers an attractive option. Additionally, our algorithm significantly reduces the computational resource and time.

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

Table 3. Performance comparing

	Cameras	Single Shot	Modify Camera	Light-Field Resolution	Training Time(h)	Reconstruct Time(h)	PSNR in dB	Disparity
MIT	1	Yes	Yes	5x5x270x480	10	18	29.1	small
Ours	5(9)	Yes	No	7x7x384x512	12.5	1.85(3.52)	33.01(38.19)	Large

* The computer for training in [14] is a workstation with a 24-core Intel Xeon processor and 200GB RAM, and the computer for reconstruction is equipped with an 8-core Intel i7 and 16GB RAM. The corresponding computer for training in our experiment is a workstation with 32-core 2.6GHz Intel Xeon CPU and 32GB RAM, and the computer for reconstruction is equipped with 8-core 3.2GHz Intel(R) Xeon(R) CPU and 8GB RAM.

We also restore 7x7 light field for different scenes (the original light field data source is from [21]) using the same dictionary, parameters and computer. As Fig. 10 shows, good reconstruction quality and large disparity are also achieved successfully for these scenes, indicating a wide applicability of our technique.

Fig. 10. Reconstruction of light field in other scene.

#216623 - $15.00 USD Received 11 Jul 2014; revised 16 Sep 2014; accepted 16 Sep 2014; published 24 Sep 2014

5. Conclusion

The proposed dictionary-based light field acquisition system uses a sparse camera array and gets rid of modification on camera(s) or other optical designs. We achieved larger disparity, higher quality and reduction of computing time. Techniques using single camera with mask inevitably lead to small disparity due to the limited size of aperture and sensor. We hope the techniques discussed in this paper provide novel and useful insights for future computational photograph and lead to effective and affordable solution to light field acquisition.

6. Future work

Our current design eliminates the mask and acquires light field with larger disparity in single shot. Light field of 7x7 views can be restored from 5 cameras with satisfactory quality. We believe the camera array could be sparser leading to significantly reduction of system's cost. We plan to explore this probability further. We also noticed that expanding the use of the trained dictionary could be limited by the light field training data. We would like to expand the range of data sets used for training. The strategy of "only local-sliding in *disaster areas*" discussed in section 3.3 effectively accelerates light field restoration without sacrificing reconstruction quality, we still need to further accelerate our algorithms by proper GPU implementation.

Acknowledgments

This work has been supported in part by the National High-tech R&D Program (863 Program) of Institute of Automation, Chinese Academy of Sciences (CASIA), grant 2012AA011903. Special appreciations to the light field data source providers who made this study possible.