Dictionary Learning based Color Demosaicing for Plenoptic Cameras

Xiang Huang
Northwestern University
Evanston, IL, USA
xianghuang@gmail.com

Oliver Cossairt
Northwestern University
Evanston, IL, USA
ollie@eecs.northwestern.edu

Abstract

Recently plenoptic cameras have gained much attention, as they capture the $4D$ light field of a scene which is useful for numerous computer vision and graphics applications. Similar to traditional digital cameras, plenoptic cameras use a color filter array placed onto the image sensor so that each pixel only samples one of three primary color values. A color demosaicing algorithm is then used to generate a full-color plenoptic image, which often introduces color aliasing artifacts. In this paper, we propose a dictionary learning based demosaicing algorithm that recovers a full-color light field from a captured plenoptic image using sparse optimization. Traditional methods consider only spatial correlations between neighboring pixels on a captured plenoptic image. Our method takes advantage of both spatial and angular correlations inherent in naturally occurring light fields. We demonstrate that our method outperforms traditional color demosaicing methods by performing experiments on a wide variety of scenes.

1. Introduction

A traditional camera cannot distinguish between different rays incident on a pixel. A light field camera, on the other hand, captures the complete $4D$ set of rays propagating from scene to camera aperture. The captured $4D$ light field contains richer scene information than a traditional $2D$ image and can be used to synthesize photographs from a range of different viewpoints or refocused at different depths [8, 7, 14, 13].

A light field camera can be implemented as a planar camera array [19], a mask based camera [18], or a lenslet-array based camera [14]. Camera arrays [19] are large, expensive, and require precise synchronization. Lenslet-based "plenoptic" camera designs are currently very popular due to commercial availability from companies such as Lytro [10] and Raytrix [16]. These cameras are portable, inexpensive, and require only a single shot to capture a light field. In this paper, we use a Lytro plenoptic camera to capture and process light fields. However, the methods presented in this paper can also be extended to other light field camera designs.

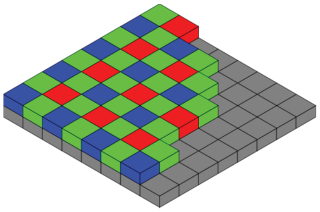


Figure 1. A Bayer color filter used to multiplex color information onto a 2D sensor. The filter consists of repeatable two-by-two grids of Blue-Green-Green-Red patterns. Bayer filters are used for both conventional 2D cameras, *and* plenoptic cameras.

Light field cameras typically capture colors in the same way as traditional cameras: by placing a Color Filter Array (CFA) on the sensor. For example, the Lytro camera uses a Bayer type CFA (Fig. 1) that is also commonly used in digital cameras, camcorders and scanners. The Bayer filter forces each pixel to capture only one red, green or blue color component. For traditional 2D cameras, a color demosaicing algorithm is used to restore missing spatial information, often incorporating an image prior to improve performance [12]. The Lytro camera places an array of around 300×300 microlenses over a 11 megapixel sensor that is covered with a Bayer CFA. The captured light field has an effective 300×300 spatial and 11×11 angular resolution. However the Bayer pattern behind each microlens introduces gaps in the full color light field: some rays in each color channel are not measured (see Fig. 2). A good color demosaicing algorithm is needed to recover this missing information in order to avoid a loss in resolution. Furthermore, since the loss of information is inherently 4D (i.e. missing rays, not pixels) the algorithm should model the captured signal as a 4D light field rather than a 2D image.

In this paper, we present a learning based technique for color demosaicing of light field cameras. We exploit

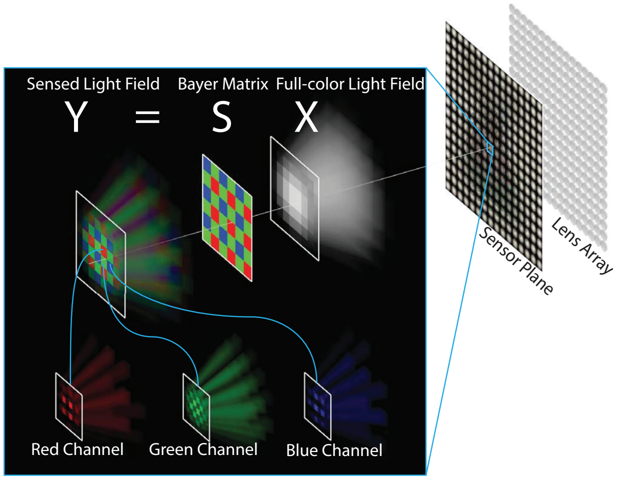


Figure 2. The Bayer filter used in a plenoptic camera causes gaps between the rays measured in each color channel. The area behind a single lenslet is zoomed in to show the effect of the Bayer filter on the captured light field. The bayer filter effectively applies a subsampling matrix S to the full-color light field X, producing the sensed light field Y. The sensed light field contains gaps: some of the rays in each of the color channel are not measured.

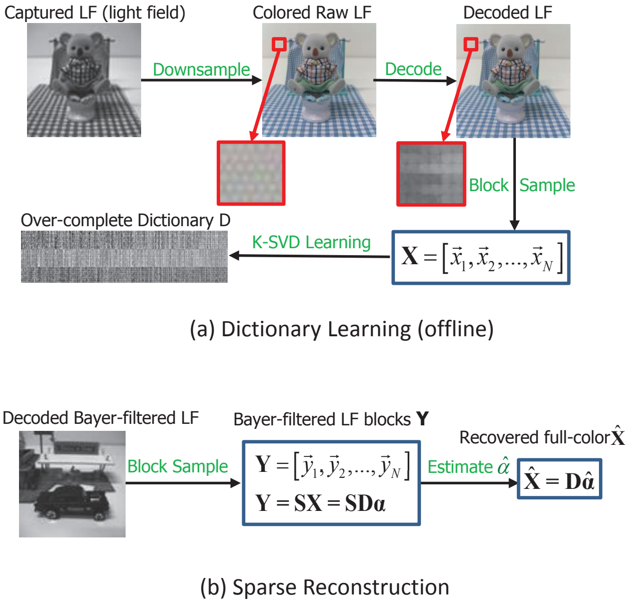
the spectral, spatial and angular correlations in naturally ocurring light fields by learning an over-complete dictionary, and reconstruct the missing colors using sparse optimization. We perform experiments on a wide variety of scenes, showing that our technique generates less artifacts and higher PSNR compared with traditional demosaicing techniques that do not incorporate a light field prior [12].

2. Previous Work

Color demosaicing algorithms for traditional cameras have been carefully studied for several decades. Those algorithms interpolate missing color values using methods that exploit spectral and spatial correlations among neighbor pixels. Examples methods include edge-directed interpolation [9], frequency-domain edge estimation [5], level-set based geometry inspired by image inpainting [6], dictionary learning [11] and gradient-corrected interpolation [12].

A significant amount of research on plenoptic cameras has been focused on modeling the calibration pipeline [4, 3] and improving image resolution [2, 17]. However, these methods use traditional demosaicing algorithms designed for $2D$ images. But plenoptic cameras capture both spatial and angular information about the scene radiance, and the best performing algorithms will model captured images as 4D light fields rather than just traditional 2D images.

Recently, Yu *et al*. [20] proposed a demosaicing algorithm for plenoptic cameras. Instead of demosaicing the raw plenoptic image, they postpone the demosaicing process until the final rendering stage of refocusing. This technique can generate less artifacts in a refocused image compared with the classical approach. However, their work is limited to the demosaicing of refocused images only. Our paper reconstructs the entire full-color light field by exploiting the spectral, spatial and angular correlations inherent in naturally occurring light fields.


Captured LF (light field)
Colored Raw LF
Decoded LF
Downsample
Decode
Block Sample
Over-complete Dictionary D
K-SVD Learning
$\mathbf{X} = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_N]$

(a) Dictionary Learning (offline)

Decoded Bayer-filtered LF
Bayer-filtered LF blocks $\mathbf{Y}$
Recovered full-color $\hat{\mathbf{X}}$
Block Sample
Estimate $\hat{\alpha}$
$\mathbf{Y} = [\vec{y}_1, \vec{y}_2, ..., \vec{y}_N]$
$\mathbf{Y} = \mathbf{SX} = \mathbf{SD\alpha}$
$\hat{\mathbf{X}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$

(b) Sparse Reconstruction

Figure 3. Overview of our approach: we learn the dictionary D from a matrix of samples X. Each column in X is a vectorized version of a block taken from a down-sampled full-color light field. The dictionary is used to reconstruct an estimate of a full-color light field $\hat{X}$ from the captured bayer-filtered light field Y. The solution is found by first finding the sparse coefficients $\hat{\alpha}$ that best represent the light field in the dictionary basis.

3. Our Approach

As shown in Fig. 3, our approach consists of two steps: a training step followed by sparse reconstruction.

In the training step, we learn all the spatial, angular and color correlations of rays in a light field from a database of raw plenoptic images captured by a Lytro camera. To generate ground truth full-color light fields, we downsample the raw Lytro images by a factor of 2, effectively reducing angular resolution by the same factor. We rectify the hexagonal-packed lenslet-array to a rectangle-packed lenslet-array to obtain a canonical plenoptic image $L(h, w)$. From the canonical plenoptic image $L(h, w)$ there is a simple mapping to the $4D$ light field $L(p, q, u, v)$, where p, q are the angular coordinates, and u, v are the spatial coordinates. Next, we sample a set of 4D blocks from the light field and lexicographically reorder to obtain a set of sample vectors. Finally, we feed the sample vectors to the K-SVD learning algorithm [1], and learn an over-complete dictionary that can be used to sparsely represent a 4D block of the full-color light field.

In the reconstruction step, we use the learned dictionary to reconstruct a full-color light field from a raw plenoptic image. We rectify the plenoptic image and divide into a set of T vectorized 4D blocks $Y = [y_i, \ldots, y_T]$. The image formation model is then given by $Y = SX$, where S is the Bayer-sensing matrix and $X = [x_i, \ldots, x_T]$ is the set of 4D light field blocks reconstructed in full-color. We apply the bayer-sensing matrix to the dictionary D and estimate a set of sparse coefficient vectors $\hat{\alpha} = [\alpha_i, \ldots, \alpha_T]$ such that $Y \approx (SD)\hat{\alpha}$. Then we reconstruct the set of full-color blocks X using linear combinations of atoms in the dictionary: $\hat{X} = D\hat{\alpha}$. Finally we reshape the matrix X into the canonical plenoptic image.

3.1. Decoding

Due to design constraints, manufacturing artifacts, and precision limitations, the microlens array of a Lytro camera is not aligned perfectly to the pixel sensor grid; the lens array pitch is a non-integer multiple of the pixel pitch, and there are unknown rotational and translational offsets. Further, the lenslet grid in a Lytro camera is hexagonally packed and must be rectified. We slightly modified the method proposed by Dansereau *et al.* [4] to decode a raw Lytro image into a canonical plenoptic image. Note that for our training set, we do not apply the demosaicing step used by Dansereau *et al.* [4], but rather down-sample to obtain the full color ground truth light field.

The canonical plenoptic image $L(h, w), h \in \{1, \ldots P \cdot U\}, w \in \{1, \ldots Q \cdot V\}$, is just a $2D$ representation of the $4D$ light field. The 4D light field $L(p, q, u, v)$ measures the rays that passes through the lenslet (u, v) falling on to the *relative* pixel (p, q) within this lenslet, where indices $p \in \{1, ..., P\}, q \in \{1, \ldots Q\}, u \in \{1, ..., U\}, v \in \{1, \ldots V\}$. The mapping between the canonical plenoptic image $L(h, w)$ and the light field $L(p, q, u, v)$ is expressed by the equations $h = p + (u-1)U$ and $w = q + (v-1)V$.

3.2. Block Sampling

We are interested in finding the sparse representation of a light field, i.e., finding a dictionary such that any light field can be described as a sparse linear combination of the atoms in that dictionary. Since a single captured light field consists of around around 10 million measurements, it is impractical to find a dictionary capable of representing such a large light field in its entirety. Instead, we decompose each captured light field in to smaller blocks. To maximally take advantage of correlations in the light field, we sample along both angular and spatial dimensions. As shown in Fig. 4, we sample a grid of $B_u \cdot B_v$ spatial positions (i.e. microlens positions), and $B_p \cdot B_q$ angular positions (i.e. pixel locations within each microlens). For training, each ground truth block is sampled from a full-color light field and has a block size of $n = 3 \cdot B_p \cdot B_q \cdot B_u \cdot B_v$. Each block is lexicographically reordered into a vector $x_i \in \Re^n$ for dictionary training. The observed signal is divided into a set of blocks represented by the vectors $y_i \in \Re^m, i \in \{1, ..., T\}$ where $m = B_p \cdot B_q \cdot B_u \cdot B_v$. Note that $n = 3m$ so that there are 3 times fewer measurements than unknowns, e.g. we want to reconstruct a 3-color light field from a Bayer-filtered one. The block size must be chosen to balance reconstruction quality and computation time, as discussed in Section 4.

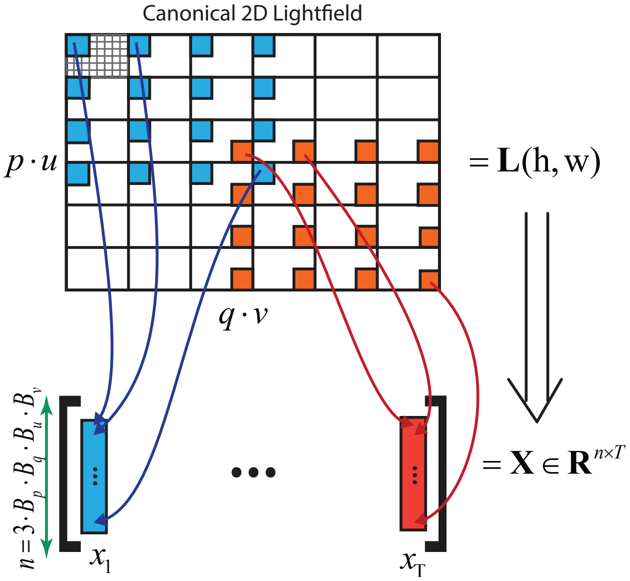

Canonical 2D Lightfield
$p \cdot u$
$= \mathbf{L}(\mathrm{h}, \mathrm{w})$
$q \cdot v$
$n = 3 \cdot B_p \cdot B_q \cdot B_u \cdot B_v$
x_1
x_T
$= \mathbf{X} \in \mathbf{R}^{n \times T}$

Figure 4. Block sampling of a canonical plenoptic image and lexicographically reordering into a vector. Here we show sampling from a block with $B_u \times B_v = 4 \times 4$ spatial samples, and $B_p \times B_q = 3 \times 3$ angular samples. With color included, the entire signal contains $3 \times 3 \times 4 \times 4 \times 3$ samples and can be represented as a $x \in \Re^{432}$ vector.

3.3. Dictionary Learning

Sparse coding is a widely prevalent tool used in image processing applications. Popular examples include $JPEG$ and $JPEG2000$ coding, which take advantage of sparsity in the discrete cosine or wavelet transform. Given a full rank dictionary matrix $D \in \Re^{n \times K}$ with K atoms, a signal $x \in \Re^n$ can be represented as a linear combination of those atoms, i.e. $x = D\alpha$. The coefficient vector $\alpha \in \Re^K$ represents the weights of atoms used to reconstruct x. For an over-complete dictionary ($K > n$, matrix D full rank), we have infinite number of solutions of α, among which the one with the fewest number of nonzero elements appeals most. We find the sparsest coefficient by solving the following sparse coding problem:

$$\min_{\alpha} \|\alpha\|_0 \quad s.t. \; \|x - D\alpha\|_2 \le \epsilon \tag{1}$$

The over-complete dictionary D can be derived analytically from a set of functions such as discrete cosine transforms or wavelets. In this paper we use a dictionary that is learned from a set of training samples. Given a set of N training samples $\mathbf{X} = [x_1, x_2, ..., x_N]$ each block-sampled from the training set of light fields, we seek the dictionary D that gives the best sparse representation for each training signal:

$$\begin{aligned} &\min_{D,\alpha_i} \sum_{i=1}^{N} \|\alpha_i\|_0 \\ &s.t. \|x_i - D\alpha_i\|_2^2 \le \epsilon \end{aligned} \tag{2}$$

We use the K-SVD algorithm [1] to optimize for the best dictionary and sparse coding of the training signals.

3.4. Sparse Reconstruction

We use sparse reconstruction to demosaic captured light fields. Demosaicing is achieved by solving the system of equations $Y = SX$, giving a solution for the set of full-color light field blocks X from the set of measured blocks Y. The Bayer sensing matrix $S \in \Re^{m \times n}$ transforms a 3-color light field into a Bayer filtered light field with 3× fewer measurements than unknowns. S is a binary $0 - 1$ matrix where the entries contain a value of 1 if and only if the corresponding color channel at that given pixel position is observed in a measured block y_i.

For our demosaicing algorithm, we apply the sensing matrix to the dictionary D and estimate a sparse coefficient matrix $\hat{\alpha}$ such that $Y \approx (SD)\hat{\alpha}$. Finally we reconstruct set of the full-color blocks X using linear combinations of atoms in the dictionary: $\hat{X} = D\hat{\alpha}$.

4. Experiments and Results

To validate the efficacy of the proposed demosaicing algorithm, we compare our method with the traditional methods of bilinear interpolation and gradient-corrected interpolation [12]. The comparison is performed on a dataset of 30 light fields captured of different scenes such as plants, fruits, flowers, toys, paintings, books as shown in Fig. 5. We split the whole dataset into a training set with 20 light fields and a testing set with 10 light fields.

For training, we randomly sample a total of approximately $20,000$ samples of block size $5 \times 5 \times 3 \times 3 \times 3$ from the 20 training light fields. From those samples, we train a dictionary D that is '2×' over-complete: it has 1350 atoms of 625 dimensional vectors. The block size directly affects the performance of our demosaicing algorithm. We experimented with different block sizes such as $3 \times 3 \times 1 \times 1 \times 3$, $5 \times 5 \times 1 \times 1 \times 3$, and $2 \times 2 \times 2 \times 2 \times 3$. We found that the block size $5 \times 5 \times 3 \times 3 \times 3$ gives the best performance within a practical training time. The dictionary is chosen to be 2× over-complete as we found using more atoms only slightly improves performance but requires longer training times. The number of samples was chosen to be around 10 times the number of atoms. We set the training residual $\epsilon = 0.01\|x\|$ since we expect a very small noise level in captured images (i.e. $SNR = 40dB = 20\log_{10}(1/0.01)$).

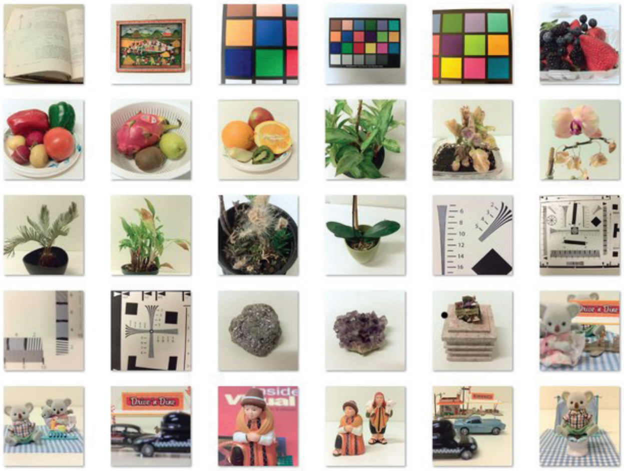


Figure 5. Our dataset of 30 light fields captured using a Lytro camera. 20 samples are used for training (i.e. learning a dictionary) and 10 samples for testing (i.e. demosaicing captured light fields).

For testing, we compared our method with bilinear interpolation and gradient-corrected interpolation [12] on a total of 10 scenes. We compute the PSNR for both the reconstructed light fields and refocused images (focused on the lenslet plane). Tab. 1 shows that our method (using block size of $5 \times 5 \times 3 \times 3 \times 3$) consistently performs better than traditional methods in all the 10 testing scenes, with an average PSNR improvement of over $5dB$. Tab. 2 shows the comparison of PSNR for refocused images. Again, our method (using block size of $5 \times 5 \times 3 \times 3 \times 3$) consistently performs better than traditional methods for all the 10 testing scenes, with an average PSNR improvement of over $4.7dB$. To show the importance of incorporating both angular and spatial correlations, we also compare results using block size of $5 \times 5 \times 3 \times 3 \times 3$ with results using block size of $5 \times 5 \times 1 \times 1 \times 3$. The former incorporates both angular and spatial correlations. It has an average improvement of $\approx 7.4dB$ (entire light field) and $\approx 7.5dB$ (refocused image) in PSNR relative to the latter which only uses spatial correlation.

We also qualitatively compare the results of our method with the gradient-corrected interpolation [12] method. Fig. 6 shows side-by-side comparison of images that are slices of light fields for a given ray angle (p, q). We can observe that our method produces significantly less visual artifacts compared to the gradient-corrected interpolation

Dataset	average	color-chart	fruit 1	fruit 2	flower	res-chart-1	stone	bear	res-chart-2	car	statue
Bilinear	30.65	32.43	3.37	29.41	33.79	31.06	32.70	29.37	28.32	27.89	30.15
Malvar [12]	31.03	32.58	32.04	29.94	34.55	31.80	33.77	29.33	28.66	28.06	29.58
Ours (55113)	28.81	31.08	29.36	27.09	31.31	28.88	30.29	27.92	26.59	26.16	29.46
Ours (55333)	36.16	37.53	37.43	35.80	39.83	35.27	39.42	35.04	33.02	32.70	35.52

Table 1. PSNR (dB) comparison of demosaicing results for 10 light field scenes. PSNR is calculated based on the reconstructed and ground truth light fields *directly*. Ours (55113) indicates a dictionary with block size of $5 \times 5 \times 1 \times 1 \times 3$. Ours (55333) indicates a dictionary with block size of $5 \times 5 \times 3 \times 3 \times 3$. We compare our method with the traditional methods of bilinear interpolation and gradient-corrected interpolation [12]. When taking both spatial and angular correlations into account (i.e. Ours 55333), our method performs $> 5dB$ greater than traditional methods.

Dataset	average	color-chart	fruit 1	fruit 2	flower	res-chart-1	stone	bear	res-chart-2	car	statue
Bilinear	39.26	40.98	39.39	37.57	41.90	39.57	40.52	38.81	37.19	37.75	39.05
Malvar [12]	41.72	42.75	42.31	40.50	45.06	42.45	43.84	40.75	40.00	39.65	40.46
Ours (55113)	38.99	41.08	39.48	37.37	41.44	39.29	40.30	38.28	37.45	36.16	39.49
Ours (55333)	46.48	47.61	47.96	46.44	50.64	46.23	49.76	45.16	43.33	43.33	45.80

Table 2. PSNR (dB) comparison of refocused images for 10 light field scenes. PSNR is calculated based on the *refocused images* generated from reconstructed and ground truth light fields. We compare our method with the traditional methods of bilinear interpolation and gradient-corrected interpolation [12]. When taking both spatial and angular correlations into account (i.e. Ours 55333), our method performs $> 4.7dB$ greater than traditional methods.

method [12].

5. Conclusion and Future Work

We have presented a learning-based color demosaicing algorithm for plenoptic cameras. By exploiting angular, spatial and spectral correlations, our algorithm performs better than traditional methods such as bilinear interpolation and gradient-corrected interpolation [12].

Our current dictionary is learned solely from a full-color light field. In the future, we are interested in exploring joint dictionary learning techniques that explicitly take into account the properties of the Bayer sensing matrix. However, the joint dictionary approach will be complicated since it requires learning a different dictionary for blocks that correspond to different portions of the bayer mask.

Our current Matlab implementation using a 2010 manufactured desktop $i7-930$ CPU takes about 3 hours for training and $40 - 50$ minutes for demosaicing a light field. We are interested in exploring faster GPU implementations of dictionary learning and reconstruction implementation such as the one in [15].

6. Acknowledgement

We thank Nathan Matsuda for help with figures. This project was supported by a Samsung GRO 2013.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K -SVD : An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. 54(11):4311–4322, 2006. 3, 4

[2] T. Bishop, S. Zanetti, and P. Favaro. Light field superresolution. *Computational Photography (ICCP), 2009 IEEE International Conference on*, 2009. 2

[3] D. Cho, M. Lee, S. Kim, and Y.-W. Tai. Modeling the Calibration Pipeline of the Lytro Camera for High Quality Light-Field Image Reconstruction. *2013 IEEE International Conference on Computer Vision*, pages 3280–3287, Dec. 2013. 2

[4] D. G. Dansereau, O. Pizarro, and S. B. Williams. Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1027–1034, June 2013. 2, 3

[5] J. Driesen and P. Scheunders. Wavelet-based color filter array demosaicking. *2004 International Conference on Image Processing, 2004. ICIP '04.*, 5:3311–3314, 2004. 2

[6] S. Ferradans, M. Bertalmío, and V. Caselles. Geometry-based demosaicking. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 18(3):665–70, Mar. 2009. 2

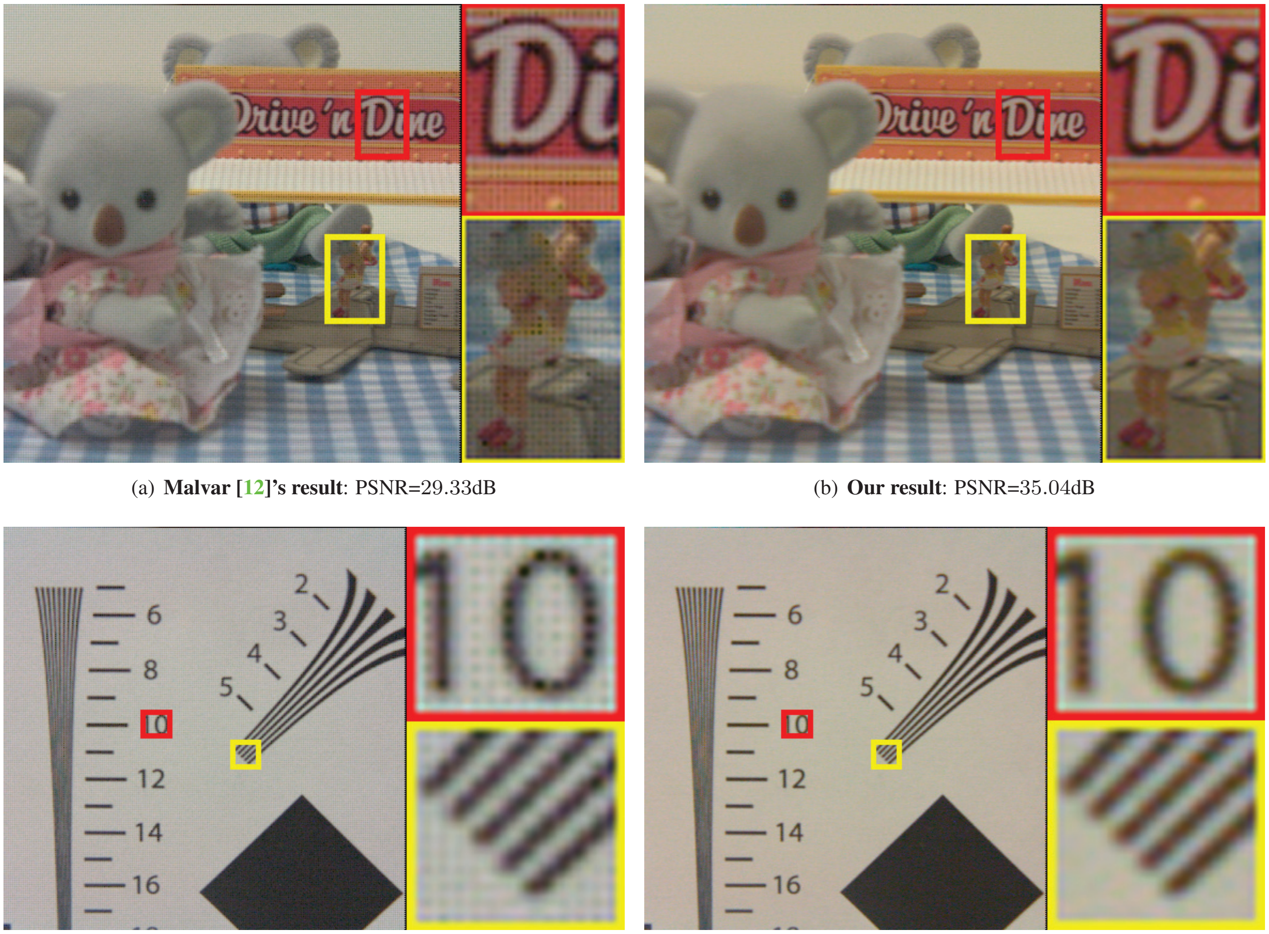
[7] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. *Proceedings of SIGGRAPH 1996, Annual Conference Series*, pages 43–54, 1996. 1

[8] M. Levoy and P. Hanrahan. Light field rendering. *Proceedings of SIGGRAPH 1996, Annual Conference Series*, pages 31–42, 1996. 1

[9] W. Lu and Y.-P. Tan. Color filter array demosaicking: new method and performance measures. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 12(10):1194–210, Jan. 2003. 2

[10] Lytro. The Lytro Camera. http://www.lytro.com. 1

[11] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE transactions on image pro-*



(a) **Malvar [12]'s result**: PSNR=29.33dB

(b) **Our result**: PSNR=35.04dB

(c) **Malvar [12]'s result**: PSNR=28.66dB

(d) **Our result**: PSNR=33.02dB

Figure 6. Comparison of demosaicing performance between our dictionary learning based algorithm (using block size of $5 \times 5 \times 3 \times 3 \times 3$) and gradient-corrected interpolation [12]. The images shown are a single view from the reconstructed light field (i.e. the set of (u, v) spatial samples for a fixed $(p, q) = (3, 3)$ angular sample). The gradient-corrected interpolation produces periodic artifacts caused by the Bayer filter. By taking into account spatial, angular, and color correlations, our method is able to reduce artifacts significantly, increasing PSNR $> 5dB$.

cessing : a publication of the IEEE Signal Processing Society, 17(1):53–69, Jan. 2008. 2

[12] H. Malvar, L.-w. He, and R. Cutler. High-quality linear interpolation for demosaicing of Bayer-patterned color images. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–485–8. IEEE, 2004. 1, 2, 4, 5, 6

[13] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics*, 32(4):1, July 2013. 1

[14] R. Ng, M. Levoy, G. Duval, M. Horowitz, and P. Hanrahan. Light Field Photography with a Hand-held Plenoptic Camera. *Main*, pages 1–11, 2005. 1

[15] R. Raina, A. Madhavan, and A. Ng. Large-scale deep unsupervised learning using graphics processors. *ICML*, 2009. 5

[16] Raytrix. The Raytrix Cameras. http://www.raytrix.de/. 1

[17] J. Stewart, J. Yu, S. Gortler, and L. McMillan. A new reconstruction filter for undersampled light fields. In *Eurographics Symposium on Rendering*, pages 1–8, 2003. 2

[18] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography. *ACM Transactions on Graphics*, 26(3):69, July 2007. 1

[19] 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. In *ACM SIGGRAPH 2005 Papers on - SIGGRAPH '05*, page 765, New York, New York, USA, 2005. ACM Press. 1

[20] Z. Yu, J. Yu, A. Lumsdaine, and T. Georgiev. An analysis of color demosaicing in plenoptic cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 901–908. IEEE, June 2012. 2