LIGHT FIELD SUPER-RESOLUTION USING INTERNAL AND EXTERNAL SIMILARITIES

Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Hanzhi Fan, Dong Liu, Feng Wu

University of Science and Technology of China

ABSTRACT

This paper presents a novel super-resolution method for light field images by jointly exploiting internal and external similarities. The internal similarity refers to the correlations that exist across the angular dimensions of the 4D light field itself, while the external similarity refers to the correlations learned from a conventional 2D image dataset. Our key observation is that the internal and external similarities are complementary to each other, and we propose a depth-adaptive fusion scheme to take advantage of both their merits. Moreover, we improve the traditional projection-based method that exploits the internal similarity, by introducing a back-projection refinement and getting rid of the dependency on camera parameters. Experimental results on a variety of light field images validate the superior performance of the proposed method.

***Index Terms*—** Light field, super-resolution, projection, learning, depth

1. INTRODUCTION

The 4D light field records light intensities not only at different spatial points but also from different angular views, which enables new applications beyond conventional 2D images, such as post-capture refocusing [1] and single-shot depth sensing [2]. Light field imaging has attracted increasing attention in recent years, mainly due to the commercialization of portable light field cameras such as Lytro and Raytrix. These light field cameras generally adopt a micro-lens-array design, which trades the spatial resolution of the sensor for the angular resolution in essence. Despite of different configurations of the micro-lens-array, the resulting light field images suffer from a considerable loss of the spatial resolution compared with the original resolution of the sensor. Therefore, light field super-resolution (SR) has induced a lot of research efforts [3, 4, 5, 6, 7].

Light field SR is quite similar to the classic multi-frame SR [8], in the sense that light field images from different angular views (i.e., sub-aperture images) have strong correlations which we name the *internal similarity*. This internal similarity provides abundant information to enhance the resolution of each sub-aperture image. Due to the rigid distribution of the micro-lens-array, the disparity between any two sub-aperture images can be characterized in a linear relationship with the unit disparity between two neighboring sub-aperture images. This property further promotes the feasibility of light field SR in comparison with multi-frame SR, as the latter has no such restriction for the disparity between different frames.

Existing light field SR methods can be roughly divided into two categories: variational-based and projection-based. Both kinds of methods exploit the internal similarity to enhance the resolution. Variational-based methods [3, 5] usually involve computationally intensive optimization. Projection-based methods [4, 6] are simple and efficient, but require the explicit depth information which is difficult to infer from the estimated disparity without knowing the camera parameters. Moreover, a main challenge for SR through internal similarity is depth discontinuity that causes occlusion in sub-aperture images. In the regions with abrupt depth changes, internal similarity no longer holds.

On the other hand, for conventional 2D images, learning-based methods have shown remarkable performance for single-image SR [9, 10, 11]. Leveraging a large image dataset, learning-based SR exploits the across-scale correlations of structural contents within natural images which we name the *external similarity*. Recently, single-image SR with deep neural networks (DNNs) [12, 13] demonstrates encouraging results. DNNs have also been applied to light field SR by directly extending the image dimension from 2D to 4D [7]. However, the networks learned from a specific light field dataset are dependent on the specific camera parameters, which may not be easily generalized to light fields obtained with different types of cameras.

In this paper, we present a novel SR method for light field images by jointly exploiting internal and external similarities. Our key observation is that the internal and external similarities are complementary to each other. Specifically, the internal similarity works well in depth continuous regions but is fragile in the regions with abrupt depth changes, while the external similarity performs best around large structures that often correspond to depth discontinuities but is less effective for fine textures that often correspond to depth continuous regions. Therefore, we propose a depth-adaptive scheme to fuse the results from projection-based light field SR and single-image SR with DNNs. Moreover, we improve the traditional projection-based method by introducing a back-projection refinement and getting rid of the dependency on camera param-

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

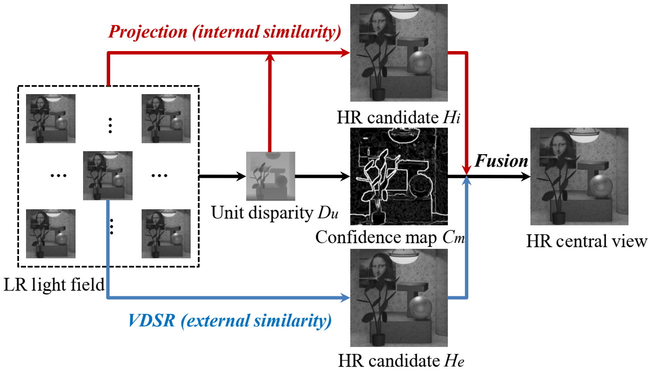

Fig. 1. Framework of our proposed method.

Table 1. PSNR (dB) results of different projection algorithms with a scaling factor of 3. PRO-N: projection without consistency checking; PRO-D: projection checking depth consistency; PRO-I: projection checking intensity consistency; PRO-IB: PRO-I with back-projection refinement.

Image	BIC	PRO-N	PRO-D	PRO-I	PRO-IB
Buddha	34.57	35.79	36.26	36.32	**36.51**
Buddha2	29.93	30.87	31.10	31.16	**31.47**
MonasRoom	34.07	33.18	34.95	35.44	**35.87**
Papillon	35.44	33.49	37.08	37.16	**37.48**
StillLife	24.48	26.38	26.86	26.87	**26.95**
Medieval	29.29	30.77	30.78	30.81	**30.87**
Average	*31.30*	*31.75*	*32.84*	*32.96*	***33.19***

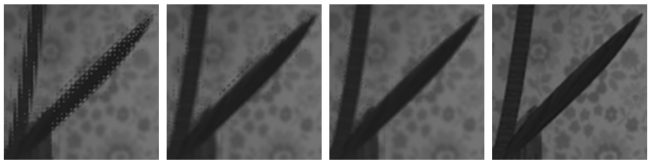
PRO-N PRO-D PRO-I Original

Fig. 2. Comparison of different projection algorithms.

eters. In this way, the advantages of internal and external similarities are jointly exploited. Experimental results on a variety of light field images validate the superior performance of the proposed method. Due to its independence of camera parameters, the proposed method can be readily generalized to light fields obtained with different types of cameras.

2. THE PROPOSED METHOD

2.1. Framework

The framework of our proposed light field SR method is shown in Fig. 1. Given a low-resolution (LR) light field, the expected output is the corresponding high-resolution (HR) light field. In this paper, we take the central view for example and SR for other angular views follows the same procedures. The first procedure is disparity estimation from the LR light field, which has been extensively studied in literature [14, 15, 16, 17]. Here we adopt the state-of-the-art method [17] with reliable performance even under occlusion. After we get the unit disparity D_u between the central view and its neighboring sub-aperture images, we can then calculate the disparity between the central view and any sub-aperture image. Assume the angular resolution of the light field is $S \times S$ (S is usually an odd value), the disparity between the central view and a sub-aperture image with angular coordinates $(u, v)(1 \leq u, v \leq S)$ can be calculated as

$$D_x(u,v) = (u - \frac{S+1}{2})D_u, D_y(u,v) = (v - \frac{S+1}{2})D_u \quad (1)$$

With the estimated disparities relative to the central view, all sub-aperture images are projected to a target image buffer, which is similar to the registration step in multi-frame SR. The next step is non-uniform interpolation, which reconstructs an HR candidate H_i for the central view image. In the proposed method, there are two improvements over the traditional projection algorithm [6]. First, we use the intensity instead of depth consistency to handle the occlusion, which gets rid of the dependence on camera parameters. Second, we impose a back-projection refinement on the reconstructed HR central view image.

On the other hand, we reconstruct another HR candidate H_e for the central view image through VDSR [13] (a single-image SR method using very deep convolutional networks). Based on the estimated unit disparity D_u which corresponds to the central view depth, we then generate a normalized confidence map C_m from the gradient of the disparity, where large values indicate abrupt depth changes. The final HR central view image is obtained by fusing the two HR candidates H_i and H_e according to the confidence map C_m. Details of the above procedures are given below.

2.2. Advanced Projection Algorithm

A main challenge for projection-based light field SR is depth discontinuity that causes occlusion in sub-aperture images. In this case, pixels that belong to the foreground in a sub-aperture image may be projected to the background in the central view image (and vice versa), which will cause severe artifacts in the regions with abrupt depth changes. To avoid the incorrect projection, the traditional algorithm checks the depth consistency to determine whether or not to project a certain pixel. Specifically, a pixel in a sub-aperture image is projected only when its depth value is consistent with that of the nearest pixel in the central view image when it is projected to the central view. The problem is, however, even we have the estimated disparities relative to the central view, converting them to depth values requires the information of camera parameters which are not easily known. Moreover, the estimated disparities may contain errors, which decreases the reliability of the depth criterion.

The independence of camera parameters will greatly promote the generality of the projection-based method. To this end, we propose to check the intensity instead of depth consistency. Specifically, a pixel in a sub-aperture image is projected only when its intensity value is consistent with that of the nearest pixel in the central view image when it is projected to the central view. Compared with the depth criterion, the intensity criterion requires no camera parameters, and will not be influenced by the disparity estimation errors.

We conduct a test on the HCI light field image dataset [18] with both information of groundtruth depth and camera parameters. The disparities required for projection are calculated from the groundtruth depth with the camera parameters, which are downsampled along with the light field images. The consistency checking during projection is then performed under two different criteria, groundtruth depth and intensity of the LR central view image. The PSNR results in Table 1 demonstrate that the intensity criterion slightly outperforms the depth criterion by an average of 0.12dB gain. (Note that the groundtruth depth is used here. The results will be degraded if the estimated disparities are used as in Table 2.) Fig. 2 shows an example for comparing different projection algorithms, from which we can see that checking the intensity consistency avoids more incorrectly projected pixels.

To further improve the projection-based method, we borrow the idea of "back-projection" from single-image SR. After pixels from all sub-aperture images are projected under the intensity criterion and the non-uniform interpolation reconstructs an HR central view image, we downsample this HR image again and calculate the difference between the downsampled image and the original LR central view image. The difference image is then upsampled through bicubic interpolation and back-projected to the HR central view image. This process is iterative and the iteration is terminated until the difference decreases to a certain value. The back-projection refinement imposes the consistency between the reconstructed HR image and the original LR image, achieving an average of 0.23dB gain in PSNR as shown in Table 1.

2.3. Combining Internal and External Similarities

The projection-based SR method exploits the internal similarity across the angular dimensions of the 4D light field itself. However, in case of depth discontinuity that causes occlusion in sub-aperture images, internal similarity no longer holds. Although checking the intensity consistency helps avoid the incorrect projection to a large extent, it also reduces the number of pixels that contribute to the resolution enhancement. In other words, the projection-based method is not effective in the regions with abrupt depth changes.

The key insight of this work is that the external similarity learned from a conventional 2D image dataset is complementary to the internal similarity. The underlying reason is that, in the natural scene, large structures often appear around depth discontinuities, and fine textures are often with continuous depth. Since learning-based SR exploits the across-scale correlations of structural contents within natural images, the external similarity performs best around large structures but is less effective for fine textures, which is complementary to the internal similarity. Fig. 3 shows two typical examples for justifying this observation, from which we can see that projection favorites textures and VDSR favorites structures.

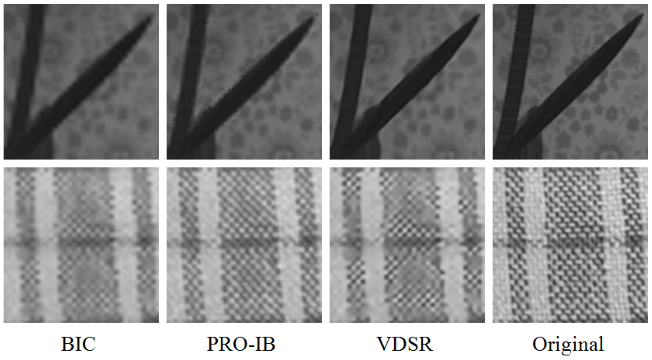

Fig. 3. Comparison of internal and external similarities.

Now that we have two HR candidates for the central view image, H_i obtained through our advanced projection algorithm and H_e through VDSR. The remaining problem is, how to fuse them into the optimal HR central view image? An intuitive way is to use the degree of depth continuity as the confidence of the two HR candidates. Here we use the estimated unit disparity D_u again. This disparity can be converted to the central view depth if the camera parameters are known. Actually, we do not need the explicit depth information, as the gradient of the disparity reflects the degree of depth continuity, where large values indicate abrupt depth changes.

Therefore, a normalized confidence map C_m is generated as follows. First, we upsample D_u to the target resolution through bicubic interpolation, based on which we calculate a gradient map. This gradient map is then linearly normalized to the interval of 0 to 1. To get rid of the influence of abnormal gradient values, gradient values larger than a certain threshold are directly set to 1. Considering the sparsity of depth discontinuity, we slightly dilate the normalized result to cover the occlusion areas adequately. Pixel values in the resulting C_m determine the weights of the two HR candidates, and the final HR central view image is generated as

$$H = C_m H_e + (1 - C_m) H_i \tag{2}$$

3. EXPERIMENTAL RESULTS

To validate the effectiveness of the proposed method, we conduct experiments on a variety of test images from the HCI

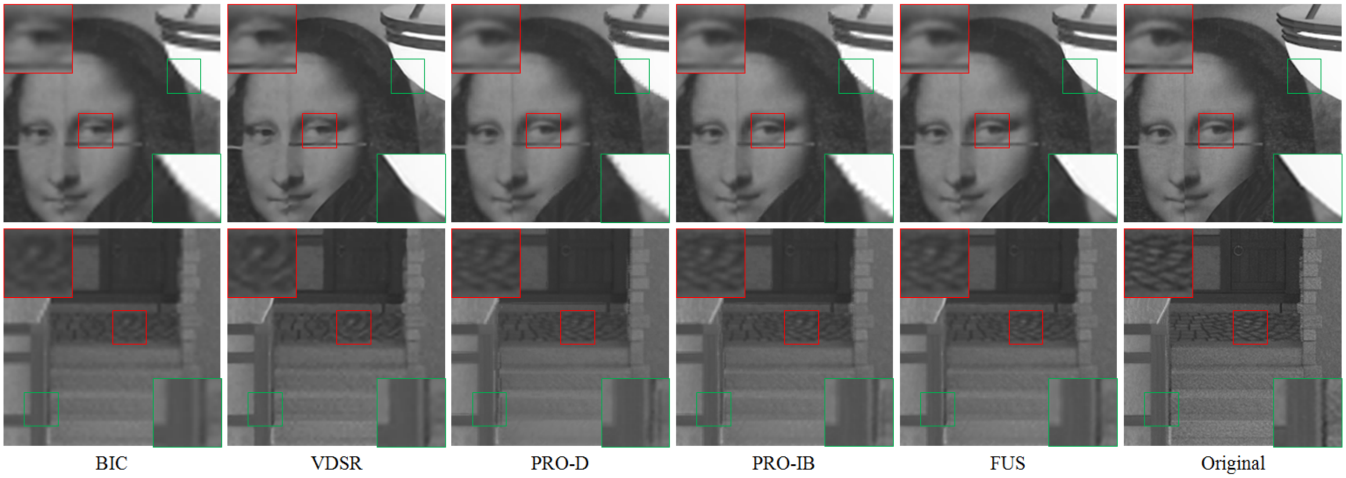

Fig. 4. Visual results of different SR methods on *MonasRoom* and *Medieval* with a scaling factor of 3.

Table 2. PSNR (dB) results of different SR methods with a scaling factor of 3 on the HCI light field dataset.

Image	BIC	VDSR	PRO-D	PRO-IB	FUS
Buddha	34.57	35.93	35.63	35.94	**36.40**
Buddha2	29.93	31.14	30.55	30.94	**31.36**
MonasRoom	34.07	36.54	35.20	35.72	**36.91**
Papillon	35.44	38.62	36.71	37.23	**38.79**
StillLife	24.48	24.74	26.49	**26.62**	26.37
Medieval	29.29	29.60	30.58	**30.70**	30.69
Average	*31.30*	*32.77*	*32.53*	*32.86*	***33.42***

light field dataset [18]. The test images are with a typical spatial resolution of 768×768 and a uniform angular resolution of 9×9. They are spatially downsampled by a scaling factor of 3 using a Gaussian filter with a standard deviation of 2. SR results of the central view through different methods are compared, including bicubic interpolation (BIC), VDSR [13], the traditional projection algorithm (PRO-D) [6], our advanced projection algorithm (PRO-IB), and our fusion method (FUS). For VDSR, a deep convolutional network is learned from a conventional 2D image dataset consisting of 500 natural images. For PRO-D, we assume the camera parameters are known so that the estimated disparity can be converted to depth. For PRO-IB, the camera parameters are no longer needed. The maximum relative errors that can be tolerated are 0.006 and 0.2 under depth and intensity criteria, respectively. For our fusion method, the threshold for distinguishing abnormal gradient values is set to 0.03, and the width of the dilation operation is 3 pixels. All these parameters are selected optimally through cross-validation.

Table 2 gives the PSNR results of different SR methods. We can see that, on average, our fusion method achieves the best performance (2.12dB gain over bicubic) and our advanced projection algorithm is the second best (1.56dB gain over bicubic). Specifically, for test images with a lot of depth discontinuities (*MonasRoom* and *Papillon*), VDSR performs much better than the projection-based methods, as the external similarity is effective in the regions with abrupt depth changes; for test images with rich fine textures (*StillLife* and *Medieval*), the projection-based methods perform much better than VDSR, as the internal similarity is effective in depth continuous regions. Not surprisingly, by jointly exploiting the two kinds of similarities in a depth-adaptive way, our fusion method generally gives the highest fidelity SR results.

Fig. 4 shows some visual results for comprising different SR methods. As can be seen, VDSR recovers sharp large structures around depth discontinuities (marked by the green rectangles) relying on the external similarity, where the projection-based methods cannot avoid artifacts as the internal similarity no longer holds. On the other hand, VDSR sometimes introduces unrealistic effects for fine textures with continuous depth (marked by the red rectangles), which reveals the deficiency of the external similarity. In contrast, the projection-based methods better preserve these structures relying on the internal similarity. In both cases, however, our fusion method generates visually pleasant results by adaptively selecting the proper similarity.

4. CONCLUSION

In this paper, we present a light field SR method by jointly exploiting internal and external similarities. The internal similarity is exploited by our advanced projection algorithm while the external similarity is exploited by single-image SR with DNNs. We further propose a depth-adaptive fusion scheme to take advantage of both their merits. Experiments validate the effectiveness of our method, which can be readily generalized to light fields obtained with different types of cameras due to its independence of camera parameters.

5. REFERENCES

[1] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report, Stanford University*, vol. 2, no. 11, pp. 1–11, 2005.

[2] Christian Perwass and Lennart Wietzke, "Single lens 3d-camera with extended depth-of-field," in *IS&T/SPIE Electronic Imaging*, 2012.

[3] Tom E Bishop, Sara Zanetti, and Paolo Favaro, "Light field superresolution," in *ICCP*, 2009.

[4] F Perez Nava and JP Luke, "Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera," in *3DTV*, 2009.

[5] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, 2014.

[6] Chia-Kai Liang and Ravi Ramamoorthi, "A light transport framework for lenslet light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, pp. 16, 2015.

[7] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *ICCVW*, 2015.

[8] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, 2003.

[9] William T Freeman, Thouis R Jones, and Egon C Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[10] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[11] Zhiwei Xiong, Dong Xu, Xiaoyan Sun, and Feng Wu, "Example-based super-resolution with soft information and decision," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1458–1465, 2013.

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.

[14] Michael W Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *ICCV*, 2013.

[15] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in *CVPR*, 2015.

[16] H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *ICCV*, 2015.

[17] T. C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *ICCV*, 2015.

[18] Sven Wanner, Stephan Meister, and Bastian Goldluecke, "Datasets and benchmarks for densely sampled 4d light fields.," in *VMV*, 2013.