TWO-STAGE CONVOLUTIONAL NEURAL NETWORK FOR LIGHT FIELD SUPER-RESOLUTION

Hanzhi Fan, Dong Liu, Zhiwei Xiong, Feng Wu*

CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System,
University of Science and Technology of China, Hefei 230027, China
fhzyzp@mail.ustc.edu.cn,{dongeliu,zwxiong,fengwu}@ustc.edu.cn

ABSTRACT

In this paper, we investigate a convolutional neural network (CNN) approach for light field (LF) super-resolution (SR). We are motivated by the assumption that image priors can be embedded into CNN, and both external and internal correlations are important in LFSR. The LF images are indeed natural images except for its angular resolution, so the external correlations help to super-resolve a single image from a collection of general images, whilst the internal correlations are essential to enhance a single view in LF with the details in the other views. Accordingly, we propose a two-stage CNN, where the two stages exploit the external and internal correlations, respectively. Moreover, to improve the generalization ability of the second-stage CNN for inter-view SR, we propose to align different views at patch level to compensate for the disparity that is essential to LFSR, thus the second stage is termed multi-patch fusion CNN. Experimental results demonstrate the superior performance of our two-stage CNN compared with the state-of-the-art CNN-based SR methods.

Index Terms— Convolutional neural network, light field, super-resolution.

1. INTRODUCTION

Light field refers to the function of light with respect to position and direction, which has been introduced into digital image processing in 1996 [1]. Recently, thanks to the portable and affordable light field cameras, especially produced by Lytro [2] and Raytrix [3], light field becomes interesting to the average consumer due to its functionalities such as depth recovery and refocusing. Since light field is virtually a 4D function, 2D for position and 2D for direction, the existing light field cameras using 2D sensor array often have to sacrifice spatial resolution to achieve angular resolution. As a result, the captured light field has limited spatial resolution, for example the Lytro Illum camera produces light field at the spatial resolution of 625×434, clearly below the current consumer expectation. How to super-resolve light field in the spatial dimension becomes an important research problem [4–8].

Super-resolution (SR) of general images has been studied for several decades, evolving from interpolation-based, reconstruction-based, to learning-based methods [9]. Recently, learning-based SR using convolutional neural network (CNN) has demonstrated remarkable progress. Dong *et al.* first proposed a CNN-based SR method known as SRCNN [10]. Later on, a very deep CNN termed VDSR is shown to outperform SRCNN significantly [11]. Indeed, CNN-based SR is essentially utilizing the correlations between general images, and learning from massive image dataset to embed the natural image priors into the learnt CNN model. To date, CNN-based methods are known to achieve the best performance for single image SR of general images.

Besides the correlations between general images, which are named *external* correlations hereafter, light field also has strong *internal* correlations, namely the high similarity between the different angular views in light field, just like that in multi-view images. The internal correlations also provide abundant information to super-resolve each view. Actually, most of the previous researches on light field SR focus on the internal correlations. They can be further categorized into variational-based [4, 6] and projection-based [5, 7] methods. However, variational-based methods require to solve optimization problems that are computationally difficult, and projection-based methods require the explicit disparity between views that is not easy to estimate if the camera parameters are not available. Moreover, Yoon *et al.* proposed a CNN-based light field SR method termed LFCNN, which directly combined neighboring views to jointly super-resolve them, to exploit the internal correlations implicitly [8]. LFCNN can enhance the spatial and angular resolution simultaneously, while we focus on spatial SR in this paper.

Motivated by the assumption that both external and internal correlations are important in light field SR, we investigate a CNN approach to exploit both correlations jointly. Specifically, in this paper we propose a two-stage CNN, where the first stage exploits external correlations by reusing a CNN for

* Corresponding author. This work was supported by the Natural Science Foundation of China (NSFC) under Grant 61331017, and by the Fundamental Research Funds for the Central Universities under Grant WK3490000001.

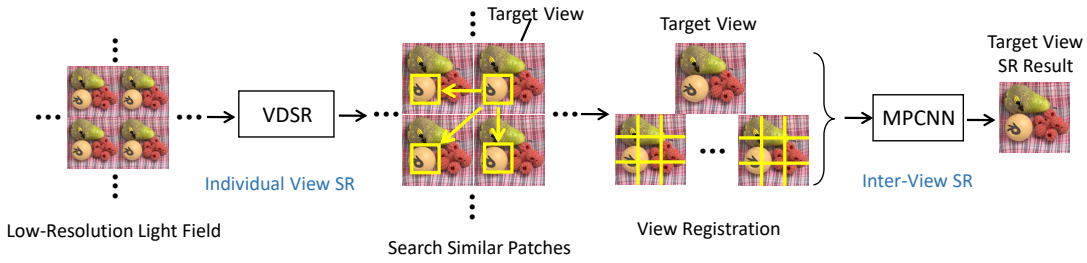

Fig. 1. Flowchart of our proposed two-stage CNN for light field super-resolution (SR). Stage 1 (VDSR) performs individual view SR, and stage 2 (MPCNN) performs inter-view SR. Registration between views is performed between the two stages.

general image SR (e.g. VDSR), and the second stage exploits internal correlations by our designed CNN to perform inter-view SR. For the second stage, different from [8], we propose to handle the disparity between views in an explicit manner. Instead of directly combining neighboring views, we perform view registration to align neighboring views before combining them. The view registration is performed at patch level, i.e. for each patch in the target view to be super-resolved, similar patches are searched from neighboring views and then constitute aligned views. Therefore, the CNN for inter-view SR is termed multi-patch fusion CNN (MPCNN).

We have performed experiments to verify the effectiveness of the proposed two-stage CNN for light field SR. Results demonstrate the superior performance achieved by our method compared with the state-of-the-art CNN-based methods. Both objective quality and subjective quality are improved by our method.

The remainder of this paper is organized as follows. Section 2 presents the details of our method. Section 3 presents experimental results, followed by conclusion in Section 4.

2. APPROACH

Fig. 1 depicts the flowchart of our proposed two-stage CNN for light field SR. Given the input of a low-resolution light field, the first-stage CNN enhances each view individually by exploiting external correlations like the single image SR of general images. The second-stage CNN, termed multi-patch fusion CNN (MPCNN), enhances the views jointly by exploiting internal correlations. Between the two stages, we perform view registration to compensate for the disparity between views. In the following subsections, we discuss the modules one by one.

2.1. First-Stage CNN

In the first stage, we propose to utilize CNN to exploit external correlations rather than internal correlations, because of the following reasons. First, external correlations can be exploited independently from light field, i.e. this stage works for any low-resolution image. Therefore, the first-stage CNN can be trained with general image dataset, not restricted to light field, which is an important advantage because the available

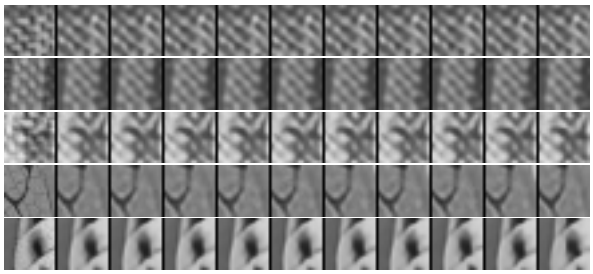

Fig. 2. Example results of searching for similar patches. The first column is the ground-truth cropped from the original target view. The second column is the corresponding patch cropped from the target view after the first stage (VDSR). The following 9 columns are the similar patches searched from neighboring views.

light field sequences are still quite scarce compared with general images. If the first stage were designed for internal correlations and the second stage for external correlations, then the second-stage CNN should be trained with only light field sequences because general images cannot fit into this scheme. Second, in this paper we propose to align views before utilizing the internal correlations, which is performed at patch level by searching similar patches. As the views have been enhanced by the first-stage CNN, we expect the patch searching can be more accurate than that performed on low-resolution views.

Since the first stage is independent from light field, any CNN for single image SR can be adopted herein. In fact, even non-CNN-based methods for single image SR are also ready to use. In this paper, we adopt VDSR [11] for its reported state-of-the-art performance, and leave the choice of other methods as future work.

2.2. View Registration

Since the second-stage CNN is to exploit internal correlations between different angular views, it is necessary to take the disparity between views into account. Since the disparity is dependent on both the scene and the camera parameters, we argue that CNN may not be efficient enough to deal with disparity natively. Therefore, in this paper we propose to per-

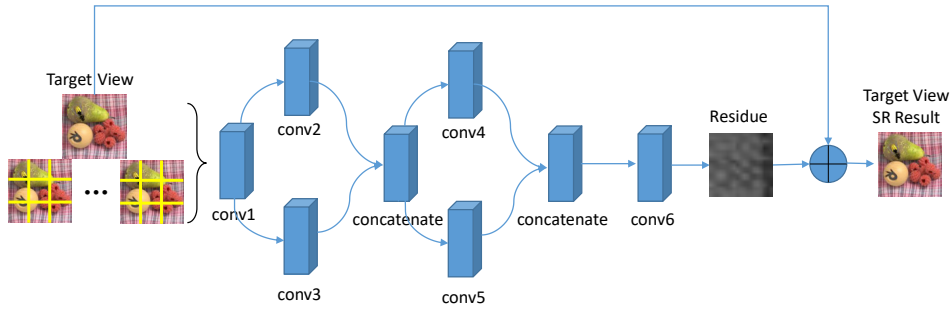

Fig. 3. The network structure of MPCNN (the second stage in Fig. 1). Detailed configurations of convolutional layers are summarized in Table 1.

Table 1. The configurations of convolutional layers of MPCNN shown in Fig. 3.

Layer	Layer1	Layer2		Layer3		Layer4
Conv. module	Conv1	Conv2	Conv3	Conv4	Conv5	Conv6
Filter size	5×5	5×5	3×3	3×3	1×1	3×3
Number of input channels	10	64	64	48	48	48
Number of output channels	64	16	32	16	32	1

form view registration between two stages.

The view registration is performed at patch level as follows. As shown in Fig. 1, take one view as the target view for example, the target view is divided into patches of the same size (19×19 in our experiments) without overlapping. For each patch, we search for similar patches from the neighboring views, where patch similarity is defined as the inverse difference between pixel values of two patches. A specific amount (9 in our experiments) of similar patches are found out and sorted by similarity in descending order. For example, Fig. 2 shows some results of the searched similar patches. Since patch search is performed at integer pixel level, the similar patches still have sub-pixel disparity, which we assume is the essential information to super-resolve the target view. Thus, we compose the similar patches to produce "aligned" views for the second-stage CNN. Specifically, the first aligned view is composed by the most similar patches to each patch of the target view, arranged at the same position. The second aligned view is composed by the second most similar patches, and so on. In total we produce 9 aligned views. Since the patches are not overlapping, the produced aligned views show kinds of "block" artifacts, but such artifacts do not affect the final SR result obviously in our experiments.

2.3. Second-Stage CNN

The second-stage CNN, termed MPCNN, is to exploit the internal correlations between views. As the views have been aligned, they are directly combined to be input into MPCNN. Thus the input of MPCNN has 10 channels (one is the target view to be super-resolved, and the other 9 are the aligned views), and the output of MPCNN is the target view SR result. The network structure and configurations of MPCNN are shown in Fig. 3 and Table 1, respectively. Our designed MPCNN has the following features. First, we adopt the residue learning strategy, i.e. the network output is supposed to approximate the difference between the original target view and the input target view (i.e. after VDSR). Learning the difference instead of the image is shown to be effective in recent work [12]. Second, the 2nd and 3rd convolutional layers of MPCNN adopt variable filter sizes, i.e. filters with two different sizes are concatenated in these layers, which improves the network ability with even less convolutional parameters [13]. Third, zero padding is used in every convolutional module to ensure the output image has the same size as the input [11]. Last but not the least, MPCNN is much smaller than the current very deep CNNs such as VDSR, as it has only 4 layers and the filter sizes are also small.

3. EXPERIMENTAL RESULTS

We use the software Caffe [14] for training our two-stage CNN as well as comparative networks. The two stages of CNN are trained sequentially. For the first stage, we follow the instructions in [11] to train the VDSR by ourselves. The training data come from the 500 natural images as used in [15]. Each image is down-sampled by a factor of 2 or 3 and then up-sampled to its original resolution to provided training samples, where the down-sampling and up-sampling filters are bicubic. Only the luminance channel is considered and reported in this paper. For the second stage, we use 9 sequences from the HCI dataset [16], known as *Buddha, Buddha2, Horse, MonasRoom, Papillon, Demo, Elephant, Statue, StillLife*, to train the MPCNN. About 569k samples are selected from these sequences, where each sample consists of a 19×19 patch cropped from a target view and 9 similar patches cropped from neighboring views. The training of MPCNN is performed at patch level similar to [11], thus the patches are not composed during training.

We then use the other 3 sequences from the HCI dataset, known as *Medieval, Screws, Motor*, and the *Tarot* sequence from the Stanford dataset[1] for testing the two-stage CNN as well as the other methods. Since the *Motor* sequence has much higher resolution (1976×1252) than the other se-

[1] http://lightfield.stanford.edu/lfs.html

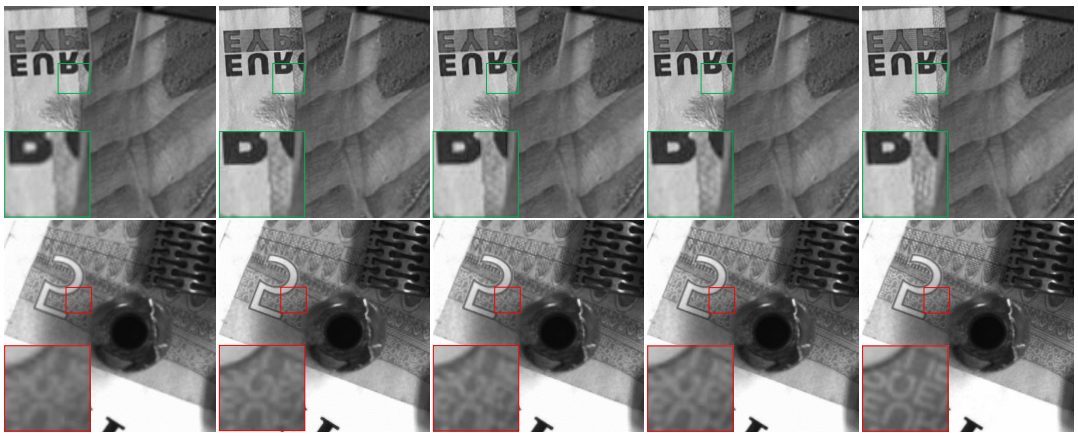

Fig. 4. Example results for comparing the visual quality of super-resolved views achieved by different methods and the ground-truth. Insets show enlarged portions of images.

Table 2. PSNR/SSIM results of different SR methods. *Motor (ds)* is the down-sampled version of *Motor*. For each sequence, **bold** indicates the best performance.

Sequence	Bicubic PSNR/SSIM	VDSR PSNR/SSIM	Inter VDSR PSNR/SSIM	MPCNN PSNR/SSIM
Medieval ×2	31.38/0.8290	32.62/0.8775	31.87/0.8604	**32.65/0.8785**
×3	29.28/0.7170	29.60/0.7545	29.52/0.7430	**29.82/0.7621**
Screws ×2	40.60/0.9833	45.40/0.9944	41.89/0.9883	**46.10/0.9949**
×3	34.44/0.9328	35.16/0.9509	35.60/0.9459	**36.64/0.9580**
Motor ×2	46.14/0.9951	46.43/0.9973	46.53/0.9958	**48.95/0.9975**
×3	38.49/0.9805	38.22/0.9848	38.65/0.9807	**39.65/0.9861**
Motor (ds)×2	34.31/0.9702	36.68/0.9829	34.87/0.9769	**38.11/0.9842**
×3	29.21/0.9143	31.25/0.9449	30.08/0.9267	**31.76/0.9488**
Tarot ×2	33.08/0.9586	36.90/0.9828	32.45/0.9431	**38.06/0.9848**
×3	27.73/0.8662	30.97/0.9375	27.10/0.8299	**31.19/0.9379**

quences, it may not benefit enough from the training data, thus we also use a down-sampled (by a factor of 2) version of *Motor* for testing. Each test sequence is entirely down-sampled by a factor of 2 or 3 and we only super-resolve its central view for example.

We compare our two-stage CNN with the bicubic up-sampling method and the CNN-based light field SR method proposed in [8]. As the method in [8] uses a shallow CNN structure, we replace it with VDSR for fair comparison[2]. Thus this method is denoted as *Inter VDSR* hereafter. The comparative results in terms of PSNR and SSIM are shown in Table 2. Reconstructed images for visual quality comparison are shown in Fig. 4.

It can be observed from Table 2 that the two stages of CNN, i.e. VDSR and MPCNN, enhance the quality of SR result gradually. For the 2× SR case, compared to bicubic, VDSR improves the PSNR of SR result by a margin of 0.29 dB to as high as 4.80 dB. MPCNN further improves the PSNR on the basis of VDSR by a margin of 0.03 dB to as high as 2.52 dB. For the *Medieval* sequence, MPCNN performs slightly better than VDSR. It shows that for this sequence, the internal correlations have not been fully exploited by the MPCNN. On the contrary, for the *Motor* sequence, VDSR improves only a little than bicubic because the input sequence already has high resolution, but MPCNN is able to improve greatly thanks to internal correlations. Thus, it demonstrates the effectiveness of exploiting both external and internal correlations in light field SR. Moreover, MPCNN also outperforms Inter VDSR consistently. It is worth noting that for the *Tarot* sequence, Inter VDSR performs even worse than bicubic, because Inter VDSR is trained on the HCI dataset but *Tarot* comes from the Stanford dataset, their inter-view correlations may be different due to camera parameters. But our MPCNN performs quite well on this sequence, because we propose to handle the disparity between views in an explicit manner, and thus improve the ability of generalization. Last but not the least, it can be observed from Fig. 4 that VDSR and MPCNN both improve the visual quality of SR result. For the regions without depth change, MPCNN performs better than VDSR because the inter-view similarity helps more in such regions.

[2]According to our empirical results, VDSR performs much better than the shallow CNN used in [8].

4. CONCLUSION

We propose a two-stage CNN-based method for light field SR. The external and internal correlations have been exploited in the two stages, respectively. Trained with a general image set, the first-stage CNN is capable in enhancing the views individually. And the second-stage CNN further enhances the target view from the information of neighboring views. Moreover, we propose to perform view registration between the two stages to handle the disparity explicitly. Experimental results show that the proposed method outperforms the state-of-the-art CNN-based methods. In the future, we plan to investigate fractional-pixel-level patch search to further reduce the interference of disparity.

5. REFERENCES

[1] Marc Levoy and Pat Hanrahan, "Light field rendering," in *Proc. SIGGRAPH*, 1996, pp. 31–42.

[2] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," Tech. Rep. CTSR 2005-02, Stanford University, 2005.

[3] Christian Perwass and Lennart Wietzke, "Single lens 3D-camera with extended depth-of-field," in *Proc. SPIE*, 2012, vol. 8291, p. 829108.

[4] Tom E Bishop, Sara Zanetti, and Paolo Favaro, "Light field superresolution," in *IEEE International Conference on Computational Photography*, 2009, pp. 1–9.

[5] F Perez Nava and JP Luke, "Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera," in *3DTV Conference*, 2009, pp. 1–4.

[6] Sven Wanner and Bastian Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, 2014.

[7] Chia-Kai Liang and Ravi Ramamoorthi, "A light transport framework for lenslet light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, pp. 16, 2015.

[8] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *IEEE International Conference on Computer Vision Workshops*, 2015, pp. 24–32.

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

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*. Springer, 2014, pp. 184–199.

[11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*. ACM, 2014, pp. 675–678.

[15] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Compression artifacts reduction by a deep convolutional network," in *ICCV*, 2015, pp. 576–584.

[16] Sven Wanner, Stephan Meister, and Bastian Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *VMV*, 2013, pp. 225–226.