Fast Light Field Reconstruction Using Convolutional Neural Network to Double Angular Resolution

Ahmed Salem, Hatem Ibrahim, Hyun Soo Kang
College of Electrical & Computer Engineering,
Chungbuk National University,
Cheongju, South Korea
ahmeddiefy@chungbuk.ac.kr, hatem@chungbuk.ac.kr, hskang@cbnu.ac.kr

Abstract—Light field imaging is the best given the amount of information it provides when compared to conventional photography, it captures angular and spatial information from all directions. Despite that, its limited resolution poses great difficulty in the use of these enormous capabilities. In this paper, we tried to lessen the impact of this drawback by using a deep-learning algorithm. We adopted the idea of dividing the process into disparity estimation and color prediction. Our system was trained to double the angular resolution fast and accurately. Experimental results demonstrate that our system can reconstruct high-quality images faster than the state-of-the-art techniques.

Keywords—Light field, view synthesis, disparity estimation, convolutional neural network.

I. INTRODUCTION

In contrast to conventional photography, the light field provides information about 3D space as it collects arrays from each direction separately. The plenoptic function represents the intensity of observed light from every direction and position. It was described using seven dimensions by Adelson and Bergen [1]. Later on, this representation was simplified using four dimensions only by Levoy et al. [2]. There are different ways to capture the light field [3], [4], and among them, micro-lens array (MLA) devices considered to provide a practical solution. Raytix adopted MLA and used in their commercial light field camera [5]. By using MLA cameras, we can avoid the disadvantages of camera arrays, which is bulky and expensive. Despite this, MLA cameras still suffer from limited sensor resolution, which affects its capability to capture dense images in the spatial and angular domains.

Due to the enormous amount of information it captures, the light field provides new applications and capabilities, such as synthetic aperture imaging [6], and segmentation and matting [7]. These enormous capabilities, coupled with the need to solve the problem of the limited sensor resolution, have inspired many researches to reconstruct light field images [8], [9], [11], [12]. Although conventional approaches provided acceptable solutions, using deep learning approaches offered remarkably better solutions. Of the approaches that used deep learning methods, Wu et al. [11] proposed to use “blur-deblur” on EPIs, although this model shows superior performance, it suffers from being time-consuming, and it did not make enough use of light field data. Kalantari et al. [12] proposed to use two separate networks for disparity estimation and color prediction, while these two networks were trained at the same time to minimize the reconstruction error. But this approach suffers from using too much time in addition to its failure in the reconstruction of some challenging scenes.

In this paper, we propose to use a deep learning framework to reconstruct light filed images. Our model was trained to double the angular resolution by reconstructing 9 x 9 output views using 5 x 5 input views. We modified the model provided by Kalantari et al. [12] to build our model. We adopted same approach of diving the process into disparity estimation and color prediction. However, we used a lighter network as well as we replaced conventional interpolation for input features extraction by a convolutional operation to speed up the reconstruction process. Our system takes 0.137% of the time used by Kalantari’s to reconstruct output views. Finally, we propose a new model to reconstruct output views very fast with high quality for real-time light field applications.

The paper is organized as follows. We present the related work in Section II. We explain the proposed method in Section III, and training on Section IV. The experimental results are presented in Section V, and the paper is concluded in Section VI.

II. RELATED WORK

In order to make the most efficient use of capabilities provided by MLA light field cameras, many solutions were proposed to reduce the effect of its limited resolution by increasing its spatial and angular resolution [8], [9], [11], [12]. Some methods proposed conventional approaches without using deep learning. Wanner and Goldluecke introduced a variational LFSR framework to enhance angular and spatial resolution using an estimated depth map [9]. Mitra et al. used a Gaussian mixture model for a patch-based approach [13], however, this approach was vulnerable to low-quality images captured by commercial cameras. Pujades et al. proposed to optimize a novel cost function using a Bayesian formulation and their approach provided better-estimated depth. Zhang et al. proposed to use Layered patch-based synthesis [14]. However, their approach shows inspiring results in some applications, but it cannot reconstruct some challenging scenes.
Many approaches are based on deep learning methods inspired by its success in different applications. Yoon et al. proposed to enhance spatial and angular resolution using CNN [15]. Kalantari et al. proposed to use two separate networks for disparity estimation and color prediction, while these two networks were trained at the same time to minimize the reconstruction error. But this approach suffers from using too much time in addition to its failure in the reconstruction of some challenging scenes [12]. Gupta et al. proposed to reconstruct light field using a single coded image [16]. Their network consists of an autoencoder and 4D CNN in parallel. But it suffers from consuming much time. Wu et al. [11] proposed to use “blur-deblur” on EPIs, although this model shows superior performance, it suffers from being time-consuming, and it did not make enough use of light field data.

III. PROPOSED METHOD

Our network is built by modifying Kalantari’s network [12]. We are following the same approach by dividing the process into disparity estimation and color prediction. 5 x 5 input views are used to generate 9 x 9 output views where every four images are processed independently to reconstruct images in between as shown in Fig. 1. First network: disparity estimation, estimates the disparity between each four input images and output images. The input for this network is 100-channel vector consists of shifted versions of input images. Each image is shifted 25 times around the position of output image using a predefined range of disparities [-8, 8], as shown by (1).

\[ I_{\text{shifted}}(S) = I(S + \text{diff} \times D) \]  

Where \( S \) is the position of each pixel, \( \text{diff} \) represents the difference in position between input and output images, and \( D \) is the predefined range of disparities [-8, -7.33, -6.67, ..., 8]. In our network, we calculated the input feature using convolution operation to achieve the required shift, where the coefficients of the convolution filters used were designed using bicubic interpolation equations. By using these bicubic convolution filters instead of conventional methods as used by Kalantari, a huge reduction in reconstruction time has been greatly reduced.

Second network: color prediction, reconstructs output image using 15-channel vectors consisting of the generated disparity, the position of the output image, and the four input images warped to the position of output image using the generated disparity. Where color prediction network learns the relation between warped input images and output image while training to achieve the best possible quality.

IV. TRAINING

Our model was trained to minimize L2 error between reconstructed and ground truth images as shown by (2).

\[ E = \sum_{\text{RGB}} (I_{\text{rec}} - I_0)^2 \]  

Where summation is over red, green, and blue channels, \( I_{\text{rec}} \) and \( I_0 \) are the reconstructed and ground truth images respectively. By applying the standard backpropagation algorithm based on gradient descent [17], the gradients are calculated at every iteration to be used to update the weights of the network while training. Where the gradients for disparity estimation and color prediction networks can be easily calculated while the gradients for the warping function are calculated numerically. By using the chain rule, gradients can flow backward through the whole system. The dataset used for training and testing are obtained from [12], [18]. The spatial and the angular resolution of input images are 541 x 376 and 14 x 14, respectively, from which, 5 x 5 images are used as input images to reconstruct 9 x 9 output images.

We used mini patches of size 10 with 60 x 60 for every patch to train our system. Xavier approach was used for weights initialization [19], while the network was trained using Adam algorithm with the default parameters [20]. We started the training process with a learning rate of 0.0001 and continued to reduce it during the training until we finished at a rate of 0.000001. We used MATLAB and MatConvNet to write our code [21]. The disparity estimation and color prediction networks almost have the same architecture as shown in Fig. 2. Each network consists of four convolution layers, where the first three layers are followed by RELU. For disparity network: the first layer has 50 filters with a size of 100 x 7 x 7, the second layer has 50 filters with a size of 50 x 5 x 5, the third layer has 50 filters with a size of 50 x 3 x 3, and the last layer has one filter with a size of 50 x 1 x 1. We used zero padding to maintain the same size for input and output images.

V. EXPERIMENTAL RESULTS

The main purpose of our design is to reduce reconstruction time as much as possible while acquiring satisfactory quality. We evaluate our model compared to others based on reconstruction time in seconds and reconstruction quality in PSNR. At first, we compare our model with Kalantari’s et al. [12], then we compare our model with Wu’s et al. model [11].

We compared our model with Kalantari’s using five test scenes where the comparison is based on Reconstruction time and quality. Fig. 3 presents a visual comparison between the ground truth and the reconstructed output images. As shown, our model outperformed Kalantari’s model regarding reconstruction time, where our model can reconstruct each output image in an average time of 0.763 seconds which represents 0.137% of the time needed by Kalantari’s model (It takes 5.56 seconds to reconstruct one output image).

We compared our model with Wu’s using the same five test scenes where the comparison is based on Reconstruction time and quality. Our model could provide average PSNR of 38.762
Fig. 2. Overview of our system. At first, input images are convolved by the bicubic filters to generate a 100-channel vector. Then, this vector is fed to the disparity estimation network to generate the disparity. The color estimation network generates the output image at the given position. Where the red block indicates a convolution filter followed by a RELU while the black box is a convolution filter only.

Fig. 3. Visual comparison of the light field reconstructed images using our approach compared with Kalantari’s et al. [12]. We used five test scenes provided by Kalantari to compare the reconstructed results. Where PSNR and SSIM are written below the images.
while the average PSNR for Wu’s model is 38.8. However, our model outperformed Wu’s model regarding reconstruction time, where our model can reconstruct each output image in an average time of 0.763 seconds which represents 0.157% of the time needed by Wu’s model (It takes 4.87 seconds to reconstruct one output image). Therefore, our system is more suitable to be used for real-time applications which require light field images reconstruction while maintaining satisfying quality.

VI. CONCLUSION

In this paper, we have presented a deep learning algorithm to reconstruct light field images. Our network consists of, disparity and color estimation networks. While the first network estimates disparity between input views and novel view. The second network uses the generated disparity to reconstruct the final images. By using the bicubic interpolation as an elementary stage to extract input features, the reconstruct time was reduced dramatically. This makes our design suitable to be used in real-time applications. Our system can achieve the minimum reconstruction time with satisfactory quality. Finally, we managed to design a high-speed model with a light network architecture, which would not need many resources to be used.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A3B03034604).

REFERENCES