DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-376

Learning based hole filling method using deep convolutional neural network for view synthesis

Heoun-taek Lim, Hak Gu Kim, and Yong Man Ro[1]; IVY Lab, Korea Advanced Institute of Science and Technology (KAIST); Republic of Korea

Abstract

In this paper, we propose a novel hole filling method in view synthesis by using deep convolutional neural networks (DCNN). The hole filling networks are learned by end-to-end mapping between hole regions and ground truth images. Hole regions are initially filled with background information. Subsequently hole filling networks restore high quality of hole filling results. The proposed hole filling networks consist of three layers, which are patch and feature extraction layer, non-linear mapping layer, and restoration layer. Experimental results demonstrate that the proposed DCNN-based hole filling method is able to significantly improve hole filling performance, compared to conventional hole filling methods. Furthermore, responses of filters learned by proposed DCNN show that the proposed hole filling framework could provide visually plausible image structures and textures to hole regions.

Introduction

In recent years, multi-view imaging systems [1], [2] such as autostereoscopic displays and free viewpoint television have drawn an attention from industry and customers. Multi-view imaging systems could provide an enhanced viewing experience by presenting different multiple perspectives of the same scene. To provide slightly different perspectives at different viewpoints, they require a large number of views. However, it is often difficult to capture, deliver, and present a great number of views at once because there are some limitations such as multi-camera capture system, network bandwidth, and display capability [2]. View synthesis techniques that generate many additional views at different virtual viewpoints could be useful in multi-view imaging systems [3]-[5].

Depth image based rendering (DIBR) is one of the widely used view synthesis techniques. In DIBR, 3D warping and hole filling are key parts to synthesize virtual views at a virtual viewpoint [6], [7]. 3D warping process is to map a given reference view to a desired virtual viewpoint using associated depth map. In this process, hole regions could be exposed in the warped view [8]-[10]. Holes are supposed to be filled with available texture information in a visually plausible manner.

In existing literatures, exemplar-based inpainting method was deployed in view synthesis in order to fill hole regions [11]. A greedy way of filling hole regions was employed, which copied the best matching local patch and pasted it to hole region. The greedy way could cause undesired visual inconsistency because the consistency between hole region and its neighboring regions did not be considered [12]. To overcome the limitation of greedy approaches, global optimization-based methods have been proposed [13]-[15]. In the global optimization framework, it is possible to provide more visually plausible results by taking into account the harmony between hole region and its neighboring regions. However, a heavy computational cost is required to solve the global optimization problem. In addition, when there is no appropriate visual information in the source region (i.e., non-hole region), most existing hole filling methods could cause visually inconsistent results that are different from the original structures.

In recent years, deep learning has succeeded in image understanding tasks such as object recognition and classification [16], [17]. Further, image de-noising and inpainting using deep learning have shown promising results [18]. These methods have shown successful results in small size regions such as super-imposed text or scratches [18], [19].

In this paper, we propose a learning based hole filling method by using deep convolutional neural networks (DCNN) in view synthesis. In particular, a new DCNN is learned for restoring the hole regions in the warped views by minimizing errors between hole region of warped view and corresponding region of ground truth image. Hole regions are initially filled with background information. Subsequently a hole filling DCNN restores initial hole filling results. The proposed DCNN for view synthesis consists of three layers, which are patch extraction and representation, non-linear mapping, and restoration. The first layer is to extract local patches and represent high dimensional feature vectors by a variety of convolutional filters. The second layer is to map the feature vectors of the first layer to another feature vectors. Thus, this layer takes into account combinations of the feature vectors extracted from initially hole-filled region and its neighboring regions. Finally, the third layer aggregate these feature vectors in order to generate a restored patch. Holes are filled by restored patches. Experimental results demonstrate that the proposed hole filling method achieves visually plausible results using convolutional filters learned by DCNN.

The rest of this paper is organized as follows. In Section 2, we present the proposed hole filling method using DCNN. Section 3 presents validation experiments that evaluate the performance of the proposed hole filling method. Finally, the conclusions are drawn in Section 4.

Proposed method

In a view synthesis, the warped view at virtual viewpoints generated by 3D warping consists of two regions, which are hole region and source region. The hole region of the warped view comes from regions occluded by foregrounds of reference view at reference viewpoint. On the other hand, the source region represents non-hole region (i.e., known regions).

The proposed hole filling method consist of initial hole filling and subsequent visually plausible hole filling by using DCNN. In the following subsections, a detailed description of the proposed hole filling method is described.

[1] Corresponding author (ymro@kaist.ac.kr)

DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-376

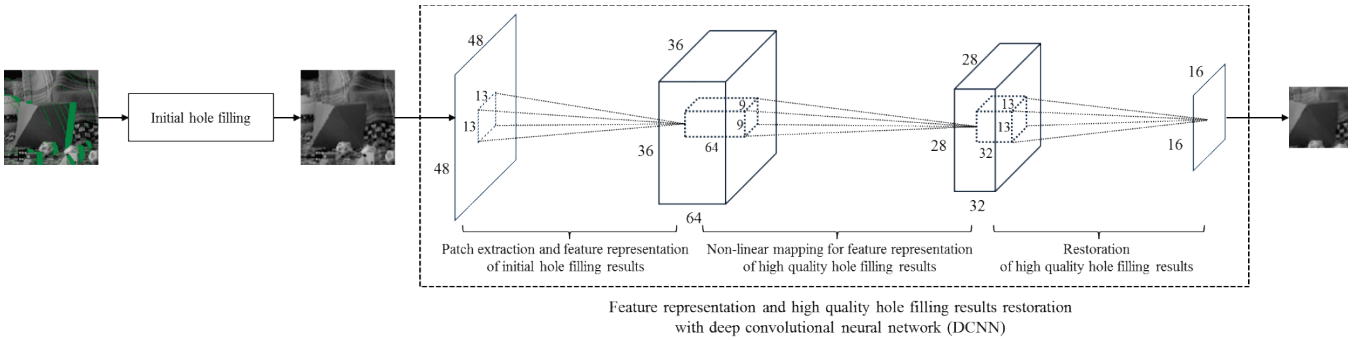

Figure 1. *Proposed hole filling method with deep convolutional neural network (DCNN) for view synthesis.*

Proposed DCNN based hole filing

The initial hole filling aims at recovering hole regions of the warped view roughly and simply. Background information of a given reference view is extracted by using disparity values. Some backgrounds are unknown (i.e., hole regions) in the warped view while they are known in reference view [10]. By referring to the corresponding pixels in reference view with disparity values, some holes in the warped view can be roughly recovered. Subsequently, remaining holes are simply filled with a conventional line-wise filling [8]. In this paper, the initial hole filling is applied in down scaled image to provide homogeneous filling in the warped view. The up scaled image of initial hole filling result is used as the input of the proposed DCNN for hole filling.

As mentioned earlier, the proposed method recovers a high quality hole filling by using DCNN from initial hole filling results. Let $\mathbf{x}^{(0)}$ and $\mathbf{y}$ denote an initial hole filling result and the ground truth virtual view, respectively. Our goal of the proposed DCNN is to restore a high quality hole filling result, which minimizes the dissimilarity with the ground truth. Figure 1 shows the proposed hole filling method with DCNN, which consists of three convolutional layers for patch extraction and feature representation, non-linear mapping for feature representation, and restoration of a high quality hole filling.

The first step of hole filling is to extract candidate patches from source region (i.e., non-hole region). Candidate patches (small rectangular regions) are supposed to contain spatial information such as textures and edges which are useful to estimate missing parts of image. When a 48 x 48 size of input patch is put in deep network, feature maps are obtained in the first convolutional layer with learned filters of size 13 x 13. They can be written as

$$\mathbf{x}_j^{(1)} = f(\mathbf{x}^{(0)} * \mathbf{W}_j^{(1)} + \mathbf{b}_j^{(1)}),\ j = 1,2,\ldots,64, \qquad (1)$$

where $\mathbf{x}^{(0)}$ is a 48 x 48 size of input sub-image (initial hole filling result). * denotes a convolution operator. $\mathbf{W}_j^{(1)}$ and $\mathbf{b}_j^{(1)}$ denote the j-th convolutional filter (size of 13 x 13) and additive bias for the j-th feature map in the first layer, respectively. $f(\cdot)$ denotes a rectified linear units (ReLU) [20].

The role of the first convolutional layer in the deep network is to extract patches and represent low-level features of the initial hole filling results. Thus, in this layer, the local structure patterns of initial hole filling results are encoded.

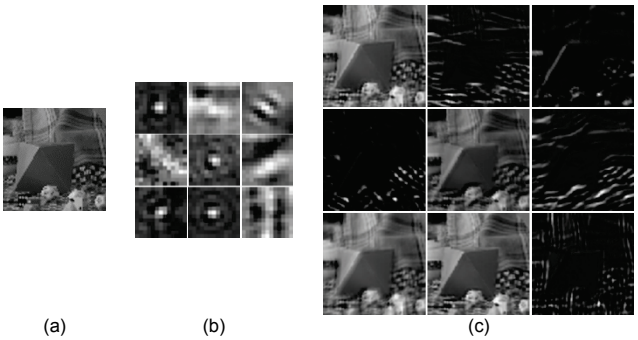

Figure 2. *Examples of learned filters and filter responses at the first layer. (a) Test sub-image (initial hole filling results). (b) Convolutional filters learned by the proposed DCNN in the first layer. The filter size was 13x13 and 9 filters were selected from 64 filters for visualization. (c) 9 feature responses obtained by learned features. Note that the black and white values represent the low and high response values, respectively.*

Figure 2 shows an example of learned filters and filter responses (i.e., feature map) in the first layer. As shown in Figure 2 (b) and (c), a variety of filters were learned for representing the structure information of initial hole filling result in the first layer.

These resulting 64 feature maps are fed to the second convolutional layer for non-linear mapping. In particular, the second layer extracts local conjunctions of features obtained in the first layer to collect high-level feature maps with neighbor information. The feature maps in the second layer can be written as

$$\mathbf{x}_j^{(2)} = f(\sum_{i=1}^{64} \mathbf{x}_i^{(1)} * \mathbf{W}_{ij}^{(2)} + \mathbf{b}_j^{(2)}),\ j = 1,2,\ldots,32, \qquad (2)$$

where $\mathbf{W}_{ij}^{(2)}$ represent the i-th convolution filters (size of 9 x 9) for the j-th feature map $\mathbf{x}_j^{(2)}$ in the second layer. $\mathbf{b}_j^{(2)}$ is a bias for the j-th feature map in the second convolutional layer.

In the second layer, the 64 feature maps including structure information of initial hole filling result mapped into 32 feature maps of a high quality hole filling result by combining the local structure patterns [21]. It can help to restore the final hole filling results in a visually plausible manner.

For visually plausible result, blending of overlapping patches using averaging filter or Poisson blending [15], [22] is employed in this paper. In the third layer, 32 feature maps are locally aggregated to generate the high quality hole filling results. They can be written as

DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-376

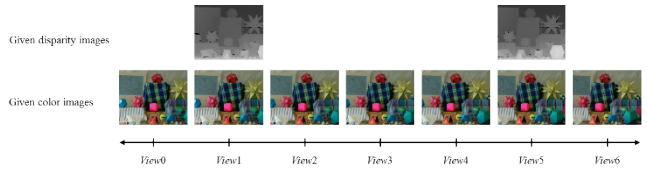

Figure 3. *View synthesis condition.*

$$\mathbf{x}^{(3)} = f(\sum_{i=1}^{32} \mathbf{x}_i^{(2)} * \mathbf{W}_i^{(3)} + \mathbf{b}^{(3)}), \quad (3)$$

where $\mathbf{x}^{(3)}$ is a 16×16 size of output sub-image, which is high quality hole filling results in our deep network. $\mathbf{W}_i^{(3)}$ is the i-th convolutional filter performing average operator and $\mathbf{b}^{(3)}$ is a bias in the third layer, respectively.

As a result, high quality hole filling results are aggregated by high-level latent features in the third layer. In this paper, the final hole filling result is obtained by replacing the hole regions of the warped view with $\mathbf{x}^{(3)}$.

Training DCNN for hole filling

In this section, we describe training procedure of the proposed DCNN, which restores high quality hole filling result from initially hole filling result. The goal of the training procedure is to estimate the parameters, which are $\mathbf{W}_j^{(1)}$, $\mathbf{W}_{ij}^{(2)}$, $\mathbf{W}_i^{(3)}$, $\mathbf{b}_j^{(1)}$, $\mathbf{b}_j^{(2)}$, and $\mathbf{b}^{(3)}$. Θ is defined as a group of parameters $\Theta=\{\mathbf{W}_j^{(1)}, \mathbf{W}_{ij}^{(2)}, \mathbf{W}_i^{(3)}, \mathbf{b}_j^{(1)}, \mathbf{b}_j^{(2)}, \mathbf{b}^{(3)}\}$. F represents the end-to-end mapping function of the proposed DCNN. The parameters are estimated by minimizing the loss function between the restored high quality hole filling results with the parameters $F(\mathbf{x};\Theta)$ and the corresponding ground truth of virtual view $\mathbf{y}$. In this paper, loss function, which is mean squared error (MSE), is used. It can be written as

$$E(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \left\| F(\mathbf{x}_i;\Theta) - \mathbf{y}_i \right\|^2, \quad (4)$$

where $E(\Theta)$ is the loss with the estimated parameters Θ. N is the number of training samples. $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i-th restored output by the proposed DCNN and the ground truth of the i-th virtual view, respectively.

In the proposed hole filling method, the DCNN is learned using stochastic gradient descent with backpropagation to minimize the loss function. By minimizing the loss function using MSE, it allows high quality hole filling results with high PSNR values ($PSNR = 10\log_{10}[255^2 / MSE]$).

Experiments and Results

To demonstrate the performance of the proposed hole filling method for view synthesis, we have performed with publicly available datasets, which are Middlebury Stereo Vision datasets [23]. Figure 3 shows view synthesis conditions in the experiment. As seen in Figure 3, the datasets consist of seven color images at different viewpoints and two disparity maps at the second and sixth viewpoints. Distance between each viewpoint is uniform. In the experiment, the color image at second viewpoint (*View*1) was warped to the fifth viewpoint (*View*4) by 3D warping in MPEG view synthesis reference software (VSRS) [10]. In the warped view, hole regions were filled by three hole filling methods, which were local greedy method [11], global optimization-based method [13], and the proposed method for performance comparison.

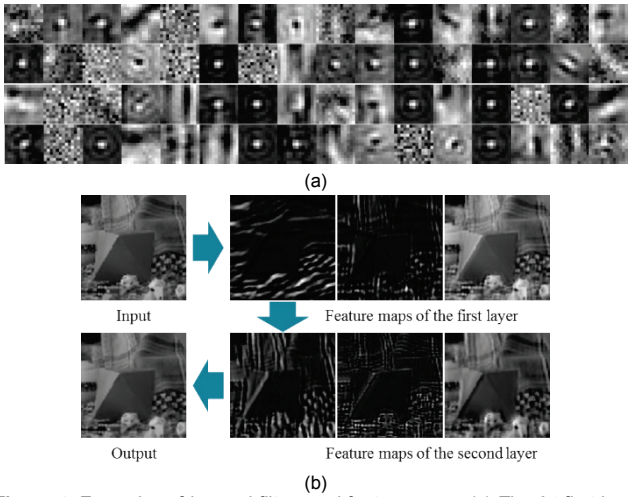

Figure 4. *Examples of learned filters and feature maps. (a) The 64 first layer filters trained on the DCNN. (b) Examples of feature maps at different layers for ""Moebius".*

In the training phase, we extracted sub-image pairs with stride 5 as DCNN learning datasets. The sub-image pairs mean that one (48×48) from initial hole filling results and the other (16×16) from corresponding ground truth images. Roughly, 680,000 sub-image pairs from 29 datasets were used in the training step. For training step in our experiment, the fourth viewpoint (*View*3) is set to the virtual viewpoint. For the test step, other 4 datasets were used, which are "Moebius", "Cloth3", "Monopoly", and "Plastic".

We implemented the proposed DCNN for hole filling using the '*Caffe*' package [24]. The training takes about two days, on GTX TITAN X GPU. In the experiment, we used only luminance channel in YCbCr color space. The chrominance channels are not used in the training step.

In the test step, we did not divide the input images to sub-images [25]. The entire initial hole filling results fed to the proposed DCNN. The proposed DCNN could provide high-resolution hole filling results. Finally, we obtained the final synthesized view by replacing the hole region of the warped view by the corresponding region in high quality hole filling result obtained by the proposed DCNN.

Figure 4 (a) shows examples of the first layer filters trained on the DCNN. As shown in Figure 4 (a), each trained filter learned its own functionality such as edge detectors at various directions and Gaussian blur. Figure 4 (b) shows examples of feature maps at each layer for "Moebius". As shown in Figure 4 (b), the feature maps on the first layer provide structure information of input such as edges. The feature maps of the second layer provide slightly different structure information.

Figure 5 and 6 show visual results of two existing hole filling methods and the proposed hole filling method for "Moebius" and "Cloth3", respectively. Figure 5 (a), (b) and Figure 6 (a), (b) show the warped image at fifth viewpoint and magnified part of (a), respectively. Figure 5 (c), (d) and Figure 6 (c), (d) show hole filling results obtained by two existing hole filling methods.

DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-376

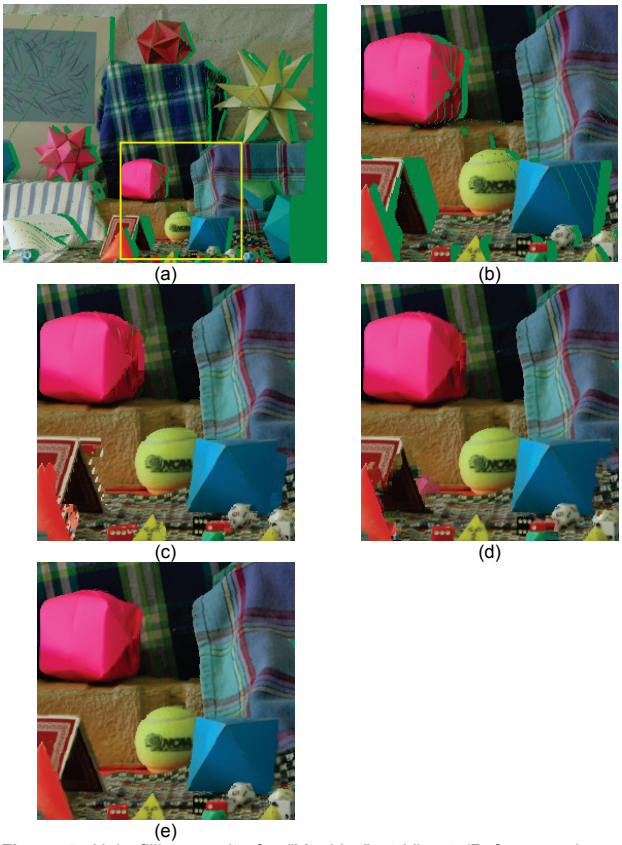

Figure 5. *Hole filling results for "Moebius" at View4 (Reference viewpoint: View1). (a) Warped view. (b) Magnified part of (a). (c) Local greedy method [11]. (d) Global optimization-based method [13]. (e) Proposed hole filling method.*

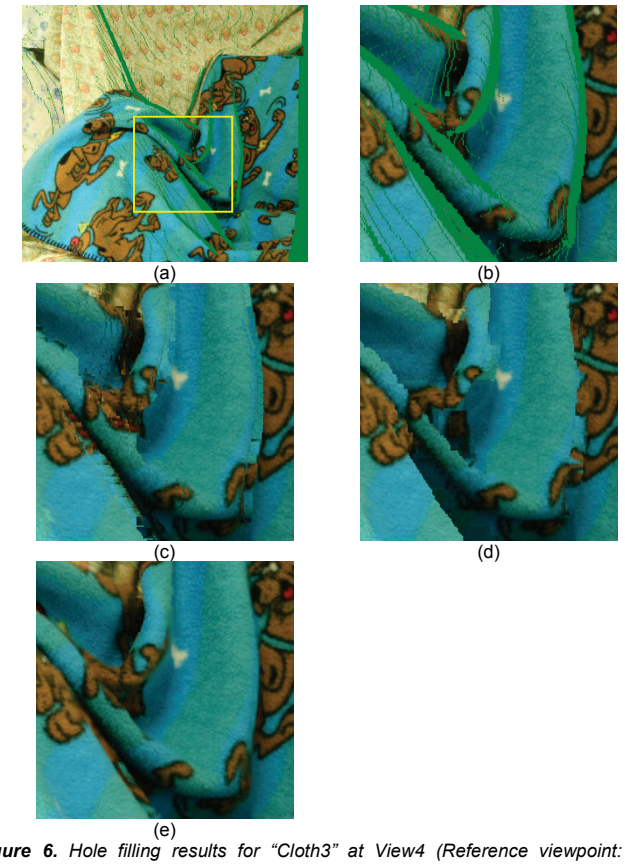

Figure 6. *Hole filling results for "Cloth3" at View4 (Reference viewpoint: View1). (a) Warped view. (b) Magnified part of (a). (c) Local greedy method [11]. (d) Global optimization-based method [13]. (e) Proposed hole filling method.*

Table 1: PSNR (dB) comparisons of hole filling results.

	Moebius	Cloth3	Monopoly	Plastic
Local greedy method [11]	27.23	27.24	28.02	33.38
Global optimization based method [13]	24.42	24.54	27.57	33.78
Proposed method	**33.04**	**33.42**	**29.93**	**35.98**

As shown in these figures, existing methods provide visually inconsistent results including structural inconsistencies. On the other hand, as shown in Figure 5 (e) and Figure 6 (e), the proposed hole filling method provide visually plausible results.

To evaluate the performance of the proposed method, we measured the quality of each hole filling result by PSNR. Table 1 shows PSNR values compared with existing hole filling methods. As illustrated in Table 1, the results obtained by the proposed method show higher PSNR values than those of existing hole filling methods. As a result, experimental results show that the proposed hole filling method can provide visually plausible results.

Conclusions

We proposed a novel hole filling method using DCNN for view synthesis. The proposed hole filling method consists of initial hole filling and three-layer DCNN structure for restoring the high quality hole filling result. Each layer of DCNN plays a key role to extract various features of the hole filling results. The proposed hole filling method with DCNN is able to significantly improve the hole filling performance, compared to conventional hole filling methods.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.2015R1A2A2A01005724).

References

[1] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. Kopylow, "A survey of 3DTV displays: Techniques and technologies," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 11, pp. 1647-1658, Nov. 2007.

[2] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, "Three-dimensional displays: a review and applications analysis," IEEE Trans. Broadcast., vol. 57, no. 2, pp. 362-371, Jun. 2011.

[3] C. Fehn, "A 3D-TV approach using depth-image-based rendering (DIBR)," in Proc. 3rd VIIP, pp. 93-104, Sep. 2003.

[4] ISO/IEC JTC1/SC29/WG11, "Reference software for depth estimation and view synthesis," Doc. M15377, Archamps, France, Apr. 2008.

DOI: 10.2352/ISSN.2470-1173.2016.14.IPMVA-376

[5] Y. J. Jeong, Y. J. Jung, and D. Park, "Depth image based rendering for multi-view generation," Journal of the Society for Information Display, vol. 18, no. 4, pp. 310-316, Apr. 2010.

[6] L. Zhang, C. Vazquez, and S. Knorr, "3D-TV content creation: Automatic 2d-to-3d video conversion," IEEE Trans. Broadcast., Jun. 2011.

[7] L. Tran, C. Pal, and T. Nguyen, "View synthesis based on conditional random fields and graph cuts," in Proc. IEEE Int'l Conf. on Image Processing (ICIP), pp. 433-436, Sep. 2010.

[8] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3D video," IEEE Trans. Multimedia, vol. 13, no. 3, pp. 453-465, Jun. 2011.

[9] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3DTV," in Proc. SPIE Stereoscopic Displays and Virtual Reality Systems, 5291, pp. 93–104, May. 2004.

[10] O. Stankiewicz, K. Wegner, M. Tanimoto, M. Domanski, "Enhanced view synthesis reference software (VSRS) for Free-viewpoint Television," ISO/IEC JTC1/SC29/WG11 MPEG2013/M31520, 2013.

[11] I. Ahn and C. Kim, "A novel depth-based virtual view synthesis method for free viewpoint video," IEEE Trans. Broadcast., vol. 59, no. 4, pp. 614-626, Dec. 2013.

[12] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," IEEE Trans. Image Process., vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[13] J. Habigt and Diepold, "Image completion for view synthesis using Markov random fields and efficient belief propagation", in Proc. IEEE Int'l Conf. on Image Processing (ICIP), pp. 2131-2134, Sept. 15-18, 2013.

[14] H. G. Kim, Y. J. Jung, S. S. Yoon, and Y. M. Ro, "Multi-view stereo image synthesis using binocular symmetry based global optimization," in Proc. SPIE, vol. 9391, pp.93910X, 2015.

[15] H. G. Kim and Y. M. Ro, "Multi-view stereoscopic video hole filling considering spatio-temporal consistency and binocular symmetry for synthesized 3D video," IEEE Trans. Circuits Syst. Video Technol., 2015 (accepted).

[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Proc. Neural Information and Processing Systems, 2012.

[17] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2015.

[18] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in Proc. Adv. NIPS, Dec. 2012.

[19] R. Köhler, C. Schuler, B. Schölkopf, S. Harmeling, "Mask-specific inpainting with deep neural networks," in Pattern Recognition, Springer, 2014.

[20] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in In Proc. ICML, 2010, pp. 131–136.

[21] Y. LeCun, J. Bengio, and G. E. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[22] N. Komodakis and G. Tziritas, "Image completion using efficient belief propagation via priority scheduling and dynamic pruning", IEEE Trans. Image Process., vol. 16, no. 11, pp. 2649–2661, Nov. 2007.

[23] Middlebury stereo datasets, available at http://vision.middlebury.edu/stereo/data/

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in ACM International Conference on Multimedia, 2014.

[25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super resolution using deep convolutional networks," IEEE TPAMI 2015.

Author Biography

Heoun-taek Lim received the B.S. degree from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2015. He is currently working toward the M.S. degree at KAIST, Daejeon, South Korea. His research interests include 3D image/video processing and free view 3D rendering.

Hak Gu Kim received the B.S. and M.S. degree from Inha University, Incheon, South Korea, in 2012 and 2014, respectively. He is currently working toward the Ph.D. degree at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include 3D image/video processing, human 3D perception, and visual quality assessment.

Yong Man Ro received Ph.D. degrees from KAIST. He was a researcher at Columbia University and a research fellow at the UC, Berkeley. He is currently a professor and the chair of signals and systems group of the school of electrical engineering in KAIST. His research interests are image processing, 3-D video processing, computer vision, visual recognition. Dr. Ro received the young investigator finalist award of ISMRM. He served as an associate editor for IEEE SPL.