Partial Plane Sweep Volume for Deep Learning Based View Synthesis

Kouta Takeuchi, Kazuki Okami, Daisuke Ochi, and Hideaki Kimata
NTT Media Intelligence Laboratories
{takeuchi.kouta,okami.kazuki,ochi.daisuke,kimata.hideaki}@lab.ntt.co.jp

Figure 1: (Left) Conventional results; synthesized images from inputs of the original plane sweep volumes (PSV) by using DeepStereo network trained by the original PSV. (Center) Our results; synthesized images from inputs of our partial plane sweep volume (PPSV) by using DeepStereo network trained by our PPSV. (Right) Ground truth; images captured by a camera at synthesized viewpoints. Although DeepStereo network trained by using the original PSV has not finished its learning sufficiently and synthesizes blurred images as shown in (Left), our PPSV enables the network to learn earlier and synthesizes high quality image as shown in (Center). Note that we used different datasets for training and prediction.

ABSTRACT

We propose a partial plane sweep volume that can be a more suitable input format for deep-learning-based view synthesis approaches. Our approach makes it possible to synthesize higher quality images with a smaller number of learning iterations, while keeping the number of depth planes.

SIGGRAPH '17 Posters, July 30 - August 03, 2017, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5015-0/17/07.
https://doi.org/10.1145/3102163.3102220

CCS CONCEPTS

• **Computer vision** → **View Synthesis**;

KEYWORDS

Novel view synthesis, plane sweep volume

ACM Reference format:
Kouta Takeuchi, Kazuki Okami, Daisuke Ochi, and Hideaki Kimata. 2017. Partial Plane Sweep Volume for Deep Learning Based View Synthesis. In *Proceedings of SIGGRAPH '17 Posters, Los Angeles, CA, USA, July 30 - August 03, 2017,* 2 pages.
https://doi.org/10.1145/3102163.3102220

1 INTRODUCTION

Creating new views from multiple pictures taken from different points has been one of the major computer vision topics for a long

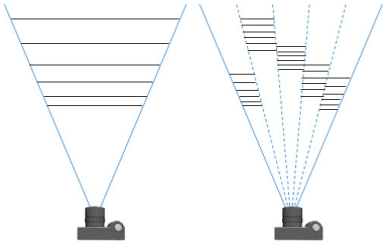

Figure 2: Schematic diagram of original and our partial plane sweep volume. (Left) Planes in original PSV. (Right) Planes in our PPSV. We individually set a range and sweep planes in each grid.

time. View synthesis is an essential technique of many attractive applications. Examples include free-viewpoint video, video frame interpolation, and video replays from the center of a soccer field. In particular, Free-D, which synthesizes replay videos of an attractive virtual camera path from several dozens of cameras in a sports stadium, has been gaining much attention in the last several years. Most view synthesis approaches from multiple images are based on a stereo matching idea. To synthesize new views, these approaches first estimate depth images by obtaining correspondences between them.This means that the estimated depth image accuracy directly affects the synthesized view image quality.

While deep learning approaches have achieved huge successes in image recognition tasks, some view synthesis approaches[Flynn et al. 2016] also adopted the deep learning idea and were able to create better quality views from smaller numbers of input views. They have shown that it is possible to train deep networks end-to-end to perform novel view syntheses. Complicated models that implicitly learn functions from networks, such as stereo matching cost or occlusion handling functions, are required to be manually designed in conventional approaches.

Most of these approaches use a set of images obtained by re-projecting images to a virtual camera from which we wish to obtain a synthesized viewpoint. This set is called a "plane sweep"[Collins 1996] and is used as input for deep learning. Using plane sweep volumes as deep network inputs makes it possible to avoid learning the relationship between camera parameters and epipolar constraints anew, and enables efficient network learning. Plane sweep volumes are usually designed so that they will encompass the volume of a scene to be synthesized. The number of planes that construct a plane sweep volume significantly affects the quality of synthesized images. Although it is desirable to set the number of planes in the volume as high as possible, increasing the number results in higher computational cost. Scenes that include both close and far objects require a huge number of planes, but this number needs to be reduced to reduce computational complexity, and this lowers the quality of synthesized images. Furthermore, the resolution of input and output images must be taken into account when considering the number of planes to be used; i.e., higher resolution requires more planes.

Accordingly, we propose a partial sweep volume that can be a more suitable input format for deep-learning-based view synthesis approaches. Our approach makes it possible to synthesize higher quality images with a smaller number of learning iterations, while keeping the number of planes. Section 2 describes the details of the partial plane sweep volume and Sec. 3 shows the obtained results.

2 OUR APPROACH

The concept of our partial plane sweep volume is to determine the position of planes in each grid in an image as shown in Fig.2. Since a sub image in a grid can be assumed to have a narrow angle of view, we can design the plane sweep volume for fitting the small depth range. To determine the position of planes, we use sparse point clouds calculated by using the structure from motion method[Wu 2013]. We describe our approach step by step below.

First, we use the structure from motion method to estimate the camera parameters of all the input images and a sparse 3D point cloud of the images' corresponding points. The "structure from motion" approach is commonly used to estimate camera parameters without any specific markers, and also used in DeepStereo.

Next, we project the point cloud to a novel view to be synthesized and get a depth map that has the sparse projected points. Then, we divide the depth image into a grid. The size of the grid is equal to the grid size defined in the following step of a deep-learning-based view synthesis approach. We select the maximum and minimum depth values from an area around each grid. From the minimum to maximum depth, for each grid, we set planes and re-project all input grid images to each plane and get the partial plane sweep volume. In contrast, conventional plane sweep volume sets the range of the planes by an entire image.

3 RESULTS

To validate our approach, we compared our partial plane sweep volume (PPSV) and the original plane sweep volume (PSV) by combining them into DeepStereo. We trained two DeepStereo networks, one by using a dataset consisting of our PPSV and the other by using the original PSV. Figure 1 (Left) and (Center) show the resulting views synthesized by these networks. Fig.1 (Left) is synthesized by using the network trained by the original PSV, Fig.1 (Center) is synthesized by using the network trained by our PPSV. In prediction steps of these networks, as inputs to these networks, Fig.1 (Left) used the original PSV, and Fig.1 (Center) used our PPSV. Note that input images of the original and our partial plane sweep volumes are same images. For training the networks, we used another 18,750 patch images for each network. Although a network trained by using the original PSV has not finished its learning sufficiently and synthesizes blurred images as shown in Fig.1 (Left), our partial plane sweep volumes enable the network to learn earlier and synthesizes high quality image as shown in Fig.1 (Center).

REFERENCES

Robert T Collins. 1996. A space-sweep approach to true multi-image matching. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*. IEEE, 358–363.

John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. 2016. DeepStereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5515–5524.

Changchang Wu. 2013. Towards linear-time incremental structure from motion. In *3DTV-Conference, 2013 International Conference on*. IEEE, 127–134.