Diminished Hand: A Diminished Reality-Based Work Area Visualization

Shohei Mori* Keio University, Japan
Momoko Maezawa† Keio University, Japan
Naoto Ienaga‡ Keio University, Japan
Hideo Saito§ Keio University, Japan

ABSTRACT

Live instructors perspective videos are useful to present intuitive visual instructions for trainees in medical and industrial settings. In such videos, the instructors hands often hide the work area. In this demo, we present a diminished hand for visualizing the work area hidden by hands by capturing the work area with multiple cameras. To achieve the diminished reality, we use a light field rendering technique, in which light rays avoid passing through penalty points set in the unstructured light fields reconstructed from the multiple viewpoint images.

Index Terms: K.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; I.4.9 [Image Processing and Computer Vision]: Applications

1 INTRODUCTION

Instructor's perspective videos and overview videos have potential to present intuitive visual instructions for trainees in medical and industrial settings. These instruction videos do not force trainees mental visual point conversion, and therefore trainees can concentrate on instructions. However, in these videos, the instructor's hands often appear and his/her work area is temporally invisible. To deal with this problem, we have proposed a diminished reality (DR) method based on light fields to visualize a hidden work area. The proposed method [3] is a light field rendering in which light rays detour points in the light fields that we refer to as penalty points. By putting a penalty point on an instructors hand, we can avoid reconstructing light rays of that area in a synthesized view. In this demo, we present an evidential system of the method named "diminished hand" using a focal points scanner (RGB-D camera) and multiple hidden ray observers (RGB cameras).

2 DIMINISHED HAND

To achieve diminished hand, we re-designed camera blending fields (CBF) of Buehler *et al*.'s method [1], which is a generalized form of free-viewpoint image generation methods using unstructured cameras. The following sections describe the re-designed CBF and view synthesis using the CBF.

2.1 Camera Pose Estimation

We use a single RGB-D camera and multiple conventional cameras that we refer to as focal surface scanner and hidden light ray observers respectively (Fig. 1). The relative poses of the scanner and the ray observers are calibrated to construct a light field. Because reference points of the following ray reconstruction (Section 2.2) is built using depth images of the scanner, all cameras are aligned regarding the scanner. First, we perform feature point matching between the scanner's depth image and the observer's images to get 3D-2D correspondences. Then, we perform bundle adjustment with the 3D-2D correspondences and camera poses calculated using the perspective-n-point problem as the initial value.

2.2 Camera Blending Field Calculation

A CBF is a map of blending weights of M ray observers (i.e., data cameras in [1]), $D_m(m = 1,2,...,M)$, in a virtual view C. In this demo, we use our work, detour light field rendering (DLFR) [3], to calculate CBFs.

3D points $\mathbf{p}_i^G$ comprise triangle meshes known as geometric proxy. These triangle meshes can be an approximation of a surface of a background or focal plane. Based on the calculated CBF, images of k most weighted data cameras are projected onto each mesh with projective texture mapping and are blended with alpha blending scheme [2]. Drawing of a triangle polygon on the estimated focal surface can be handled independently so that it can be processed at high speed by parallel processing. The proposed method uses geometry shader scheme to duplicate one triangle into three to blend them at once, while general image-based rendering performs three draw calls for one triangle to blend three vertices' weights.

2.3 Focal Points Estimation

An array of cameras can synthesize a digitally refocused image (e.g., SAP) to blur the foreground to make it virtually invisible. However, there are many constraints on camera arrangement and real-time processing. The proposed method makes similar effects using less number of cameras (e.g., four cameras). To achieve efficient refocusing, we reconstruct a background depth image at the instructor's perspective.

First, the scanner's depth map is back-projected as a 3D point cloud $\mathbf{p}'^G_i$. The 3D points are transformed to the instructor's viewpoint, thereby we obtain a depth image I_D at the instructor's perspective. Although some regions are missing or cannot be observed from the scanner, they can be filled with surface splatting [4] and the past depth frames. After initializing I_{D} at the initial frame, this depth image is updated by the weighted average with the previous depth image at each frame. In addition, hand depth map I_{H} and focal surface depth map I_{F} are separated by thresholding in the world coordinate system.

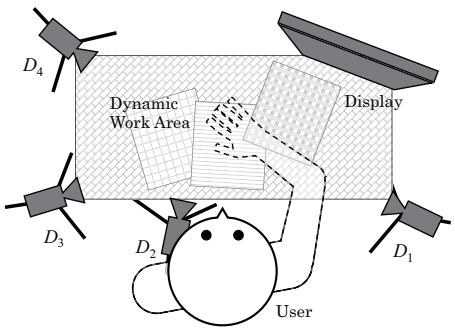

Figure 1: Illustration of our demo

*e-mail: mori@hvrl.ics.keio.ac.jp
†e-mail: momoko_maezawa@hvrl.ics.keio.ac.jp
‡e-mail: ienaga@hvrl.ics.keio.ac.jp
§e-mail: saito@hvrl.ics.keio.ac.jp

2017 IEEE Virtual Reality (VR)
March 18-22, 2017, Los Angeles, CA, USA
978-1-5090-6647-6/17/$31.00 © 2017 IEEE

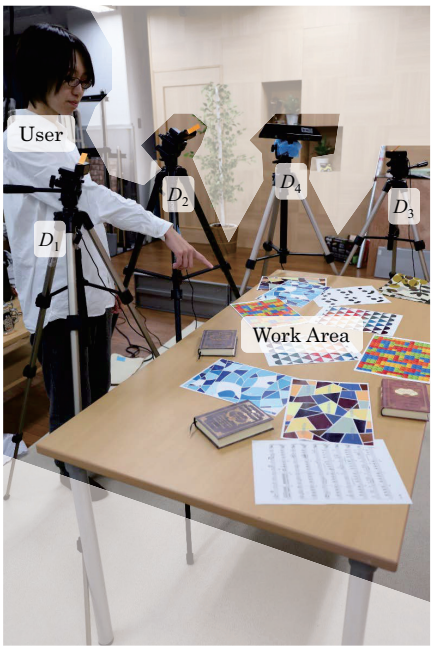

Figure 2: Experimental system setup

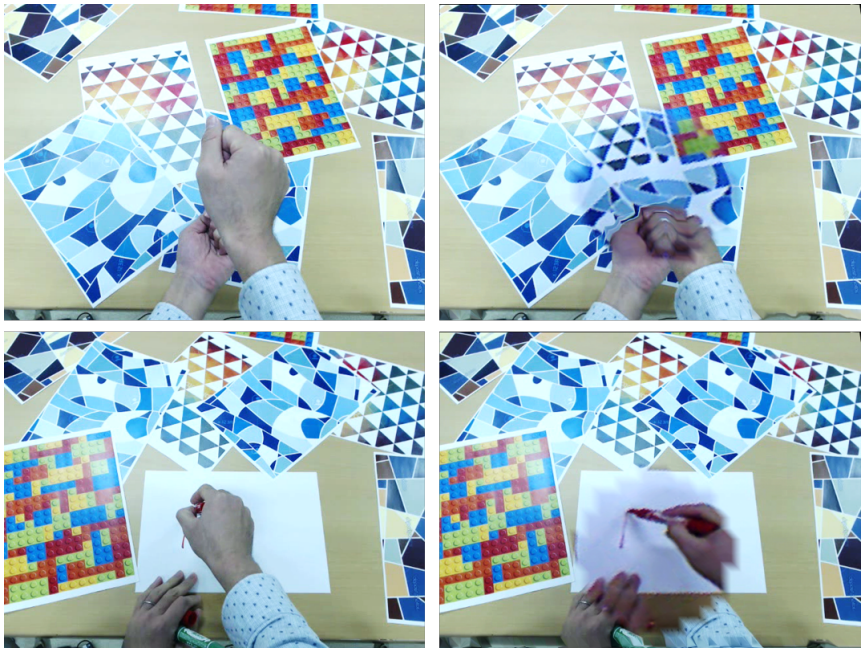

Figure 3: Example results of the experimental system. Left column shows input and right column shows diminished reality results.

2.4 Hand Detection

The area to be removed is determined using I_{H}. A 3D position corresponding to the pixel existing at the highest position among the valid pixels in I_{H} is considered as the position of a fingertip. Therefore, an arbitrary offset is given to adjust the position and a volume of interest is determined as a sphere centered on the 3D point having a certain radius.

3 Results and Discussion

3.1 Setup

We present experimental results to show that the proposed system can visually removes an undesirable hand in real time. We used a Microsoft Kinect sensor and three USB cameras (640×480 resolution) facing planar and non-planar desktop work areas (Fig. 2). We obtained intrinsic and extrinsic parameters of a virtual camera C and data cameras $D_m(m = 4)$ by bundle adjustment described in Section 2.1. Total time required for the setup was about 30 minutes.

3.2 Implementation

The proposed system worked on a Windows 10 64-bit laptop with an Intel Core i7-6567U 3.30 GHz CPU, Intel Iris Graphics 550 GPU, and 16.0 GB memory. The CBF calculation [3] and view synthesis based on the CBF (Section 2.2) is implemented on CPU and GPU respectively. The system was implemented using C++ and OpenGL shading language 3.3.

3.3 Results

Fig. 3 and a video[1] show examples of DR results. In the top row of the example results in Fig. 3, the user waved his right hand in the midair and the other hand is placed on the desktop. In the bottom result, the user draw something on a paper using his right hand and the hidden drawings are seen through the hand. In both cases, the hand automatically detected and visually removed in the view. The resulting video illustrates our system operation.

[1] Video: https://youtu.be/BrEBqynUE6Q

Average frame rates of the proposed method were about 40.1 fps and thus the proposed method works at above video rates. These results demonstrate that our system visually removes a hand in real time. Jitter and distortion which is considered to be caused by inaccurate focal surface estimation were sometimes seen due to the focal point updates at every frame. Object removal becomes challenging as the hand approaches to the background because the background is no more visible at any data cameras.

4 Conclusion

In this article, we proposed a real-time diminished reality system, which uses detour light fields to visualize occluded work areas in an instructor's perspective. The light field via unstructured multiple views is designed to avoid passing through spaces determined by penalty points. Therefore, an instructor's hand within the space is diminished in the instructor's perspective video. Results of an example setup showed that the proposed method removes a hand from an image in real time.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research (S) Grant Number 24220004 and a Grant-in-Aid from the Japan Society for the Promotion of Science Fellows Grant Number 16J05114.

References

[1] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Proc. SIGGRAPH*, pages 425–432, 2001.

[2] P. Debevec, Y. Yu, and G. Borshukov. Efficient viewdependent image-based rendering with projective texture-mapping. In *Proc. Eurographics Rendering Workshop*, pages 85–92, 1998.

[3] S. Mori, M. Maezawa, N. Ienaga, and H. Saito. Detour light field rendering for diminished reality using unstructured multiple views. In *Proc. Int. Workshop on Diminished Reality*, pages 292–293, 2016.

[4] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Surface splatting. In *Proc. SIGGRAPH*, pages 371–378, 2001.