OSTNet: Calibration Method for Optical See-Through Head-Mounted Displays via Non-Parametric Distortion Map Generation

Kiyosato Someya*
Tokyo Institute of Technology

Yuichi Hiroi
Keio University

Makoto Yamada
Kyoto University

Yuta Itoh
Tokyo Institute of Technology

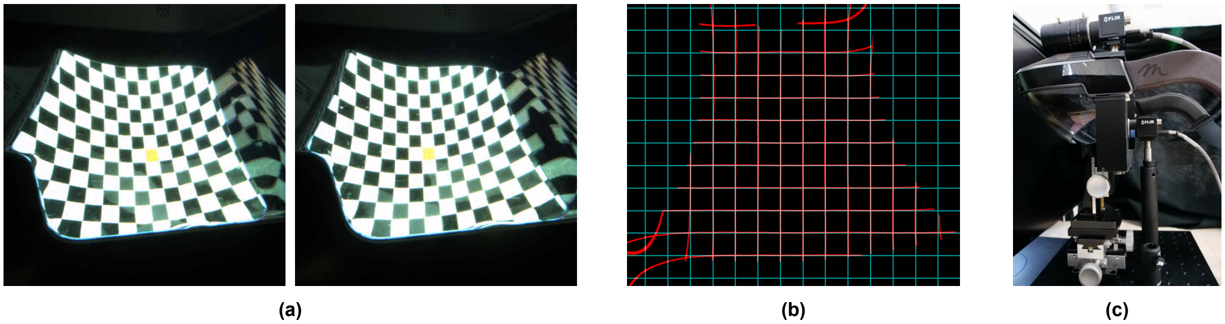
(a) (b) (c)

Figure 1: (a) Actual shots of a wide-FoV OST-HMD taken by a viewpoint camera. The distortion patterns change non-linearly for each viewpoint. (b) Visualization of a calibration result of our method calculated in the viewpoint camera image space. Red grids are calculated observation with the calibration. If calibration is perfect, the red grids should exactly align with the base green grids. (c) A side view of the actual calibration setup with an OST-HMD and cameras.

Abstract

We propose a spatial calibration method for Optical See-Through Head-Mounted Displays (OST-HMDs) having complex optical distortion such as wide field-of-view (FoV) designs. Viewpoint-dependent non-linear optical distortion makes existing spatial calibration methods either impossible to handle or difficult to compensate without intensive computation. To overcome this issue, we propose OSTNet, a non-parametric data-driven calibration method that creates a generative 2D distortion model for a given six-degree-of-freedom viewpoint pose.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality Computing methodologies—Computer graphics—Graphics systems and interfaces—Virtual reality

1 Introduction

In Augmented Reality (AR) with Optical See-Through Head-Mounted Displays (OST-HMDs), spatial calibration—spatially aligning virtual contents into the real scene at the viewpoint of the user, is a crucial requirement to maintain the realism of AR contents.

Most of the early works treat the image screen of an OST-HMD as a planar plane. However, some others report that the triangulation of the image screen gives a distorted curved plane [7]. Complex optics of OST-HMDs could distort incoming light from microdisplays and the resulting screen image may not necessarily in planar. The distortion could even vary in its shape depending on the eye position.

Several works employ more complex parametric distortion models or even non-parametric distortion mappings [1, 6]. While the non-parametric methods can handle the distortions of both virtual image and the real scene, their mapping requires ray tracing at the rendering, which might be an extra computing cost.

Klemm et al. proposed another non-parametric method where they measure distorted images from several viewpoints as reference data and apply a simple linear approximation for a given new eye position [3–5]. Their method also integrates 2D distortions into a rendering pipeline for practical use. Their dense data sampling, however, only distributes in 2D directions from the original viewpoint and does not consider the depth change of the eye position.

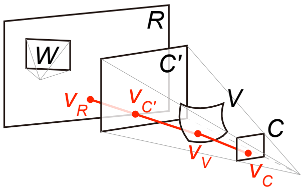

Figure 2: Coordinate systems used in this paper. C corresponds to the viewpoint camera, V the OST-HMD, C' the ideal image, R the background display, and W the world camera.

In this work, we propose a 2D distortion map approach that can handle eye position-dependent, pixel-wise distortion based on a generative model. The smooth variation of distortion brought us a data-driven approach where we learn the distortion maps as a function of eye poses, which resulted in using Variational Auto Encoder (VAE [2]): a generative deep-learning approach.

2 Method

We first define coordinate systems in the calibration problem (Fig. 3). C is the 2D coordinate system of a viewpoint camera. V is the 2D coordinate system of an OST-HMD optics where a pixel is obsevered from the viewpoint. The coordinate system V is distorted by its extraordinary optics. C' represents the 2D coordinate system of a virtual, ideal image. The virtual plane is placed on an extension of the viewpoint camera. R is the 2D coordinate system of a planar background display. W is a 3D coordinate system of a world camera, which is mounted on the OST-HMD.

Additionally, v_C, v_V, $v_{C'}$ and v_R denote the 2D points on corresponding coordinate systems. Note that $\overrightarrow{v_C v_V}$ and $\overrightarrow{v_{C'} v_R}$ are not necessarily on the same ray since the beam splitter in the OST-HMD refracts the incoming ray from outside of the OST-HMD.

For the calibration, we need to derive $M_{VC'}(v_V)$, the relationship between v_V and $v_{C'}$, from $M_{RC'}(M_{CR}(M_{VC}(v_V)))$. Thus we need

*e-mail: kiyosato@ar.c.titech.ac.jp

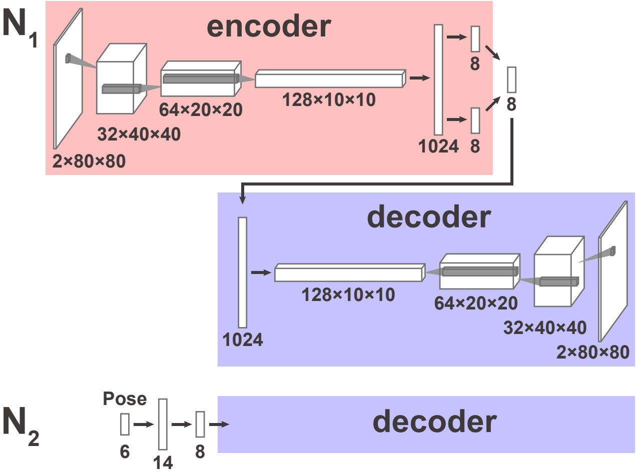

Figure 3: Network architectures used in our method. (Top): The VAE architecture to create $M_{VC'}$ from $M_{VC'}$. (Bottom): The fine-tuning model to create $M_{VC'}$ from a given viewpoint pose P, where we connect a pre-trained decoder used in the above VAE model.

$M_{RC'}$, M_{CR} and M_{VC}.

We first display a gray-code pattern on the background display R while capturing them by the viewpoint camera without placing the OST-HMD. We define that the ideal image position C' has the same view angle of C, thus these gray code patterns represent the pixel-wise correspondence of coordinates between R and C'. By using these images, we acquire the correspondence map $M_{RC'}$ as a form of a lookup table. After attaching the OST-HMD, M_{CR} and M_{VC} can be acquired in the same manner.

To obtain the view-dependent map directly from the viewpoint P, we model $M_{VC'}(P)$ as a non-linearity mapping defined by the combination of the poses and the decoder of VAE, as Fig. 3. We obtain the world coordinates of each $P \in \mathbb{R}^6$ through the world camera W.

We take the difference from the reference point of $M_{VC'}(P_i)$ look-up table and resize them to 64×64 pixel. We concatenate the table of x and y, then pad with zeros around the border of the images. As a result, we get $2 \times 80 \times 80$ tensor of inputs of the N_1. To generalize the performance of the network, we generate 100,000 samples from the training data by using mix-up [8] data augmentation.

After N_1 is trained, we concatenate a fully-connected network whose input is $\mathbb{R}^6$ viewpoint pose vector, in front of the decoder of N_1. N_2 denotes this concatenated network. When learning N_2, the decoder is kept unchanged.

During the inference, we put any viewpoints P' in N_2, then we can get the inferred correspondence map $M_{VC'}(P')$ from the output of N_2.

3 Analysis and Result

We compare the accuracy of our method with a linear method(trilinear interpolation). For each analysis, the two methods use the same datasets.

We mount the OST-HMD (Meta2) and the world camera on a composite translation stage, which moves in x-, y-, and z-direction respectively. We rigidly fixed the viewpoint camera and the background display, and we move the OST-HMD.

We get the datasets of $M_{VC'}$ and P from 173 different viewpoints distributed in 3D. The points are inside an eyebox cube of 12mm on one side. We use 125 datasets of them as training data and 48 of them as test data.

Table 1 gives the result of the quantitative analysis by showing the errors of $v_{C'}$ by OSTNet and the linear method. The pixel errors in the table represent V pixels. The angular error is the error converted to a viewing angle when the C' is viewed from the position of the viewpoint camera.

Table 1: Results of the accuracy analysis.

Method	Pixel Error (pixel)			Angular Error (arcmin)		
	Mean	Med.	Max	Mean	Med.	Max
OSTNet	2.20	1.93	25.55	9.68	7.01	103.00
Linear	2.15	1.89	11.82	8.68	6.98	52.71

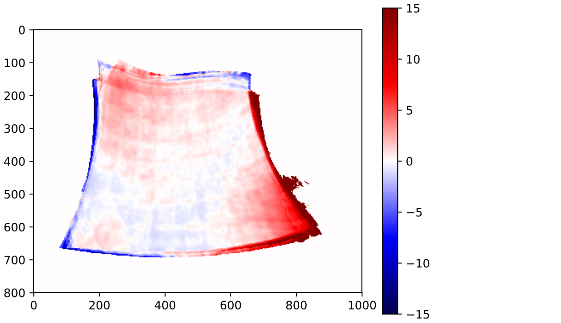

Figure 4: The difference of the error averaged at each pixel of the view point camera. The red part means the linear interpolation is better, and blue part means the OSTNet is better.

We further analyze the calibration errors in terms of the FoV of the viewpoint camera. Figure 4 shows the difference of the error averaged at each pixel of the viewpoint camera across all view poses between OSTNet and Linear interpolation. From the figure, we see that both methods have small errors near the center of the FoV and have large errors around in the peripheral of the FoV.

4 Conclusion

We presented OSTNet, a deep learning-based spatial calibration method for OST-HMDs. The result shows that our method provides an average error of about 9.68 arcmin, which is comparable to a dense linear interpolation method. We further provide a discussion on how we can further improve our method including improving the pre-training process and optimizing the network architecture.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Numbers JP17H04692, Japan; and by JST PRESTO Grant Number JPMJPR17J2, Japan.

References

[1] Y. Itoh and G. Klinker. Light-field correction for spatial calibration of optical see-through head-mounted displays. *IEEE TVCG*, 21(4):471–480, 2015.

[2] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] M. Klemm, H. Hoppe, and F. Seebacher. [poster] non-parametric camera-based calibration of optical see-through glasses for augmented reality applications. In *IEEE ISMAR*, pp. 273–274. IEEE, 2014.

[4] M. Klemm, F. Seebacher, and H. Hoppe. Non-parametric camera-based calibration of optical see-through glasses for ar applications. In *IEEE Cyberworlds*, pp. 33–40. IEEE, 2016.

[5] M. Klemm, F. Seebacher, and H. Hoppe. High accuracy pixel-wise spatial calibration of optical see-through glasses. *Computers & Graphics*, 64:51–61, 2017.

[6] S. Lee and H. Hua. A robust camera-based method for optical distortion calibration of head-mounted displays. In *IEEE VR*, pp. 27–30, 2013.

[7] C. B. Owen, J. Zhou, A. Tang, and F. Xiao. Display-relative calibration for optical see-through head-mounted displays. In *IEEE ISMAR*, pp. 70–78, 2004.

[8] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017.