Motion parallax representation for indirect augmented reality

Fumio Okura[*] Osaka University

Yuya Nishizaki NAIST

Tomokazu Sato[†] NAIST

Norihiko Kawai[†] NAIST

Naokazu Yokoya[†] NAIST

Abstract

Indirect augmented reality (IAR) presents pre-generated augmented images for achieving high-quality geometric and photometric registration between pre-captured images and virtual objects. Meanwhile, IAR causes spatial inconsistency between the real world and presented images when users move from the location where the real scene was captured. This paper describes a novel way to address the spatial inconsistency; namely, enabling viewpoint change in IAR. The key idea of this study is to employ an image-based rendering technique using pre-captured multi-view omnidirectional images to provide free-viewpoint navigation. For a pilot study, we have developed an IAR system representing a motion parallax effect using an optical-flow-based camera motion estimation.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities

1 Introduction

In augmented reality (AR) applications using mobile devices, geometric registration that accurately aligns virtual objects onto real scenes is one of the most essential requirements. As a unique approach to address the registration problems, indirect AR (IAR) has been proposed [8]. This approach captures an omnidirectional image of a target scene and superimposes virtual objects on the image in advance. The online process crops the pre-generated augmented image according to the device orientation obtained by internal sensors and presents the cropped image to users. Unlike traditional AR that utilizes real scenes captured in real-time, IAR never produces jitters between pre-captured scenes and virtual objects.

However, IAR causes inconsistency between the real world and presented images, which can be categorized into spatial and temporal ones [5]. One major issue of IAR is spatial inconsistency; i.e., traditional IAR cannot present scenes from viewpoints other than where omnidirectional images have been captured. This study proposes an IAR approach presenting scenes from viewpoints different from those for input images according to user's device poses, which allows users to move around, while preserving the advantage of jitter-free AR. To achieve this, we employ image-based rendering using omnidirectional images captured at multiple locations.

For the first step of the study, we have developed a prototype system representing relatively small viewpoint change, i.e., motion parallax. We propose an optical-flow-based estimation of relative motion of the device camera, which is robust to texture-less scenes compared with visual simultaneous localization and mapping (vSLAM)-based camera pose estimation approaches.

2 IAR with Motion Parallax Representation

The proposed approach generates IAR scenes with viewpoint changes using image-based rendering. Our prototype system does not target large viewpoint change but handles small motion of the viewpoint such as motion parallax by optical-flow-based camera motion estimation. Similarly to previous IAR systems [5, 8], our approach consists of offline and online processes.

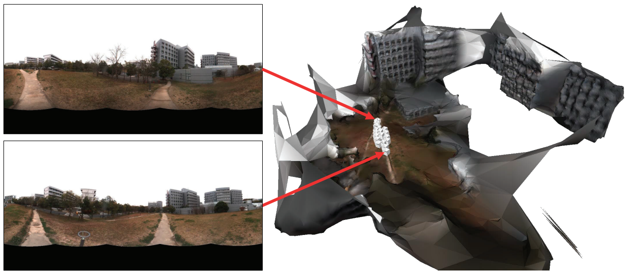

Figure 1: Examples of input omnidirectional images and a reconstructed 3D model.

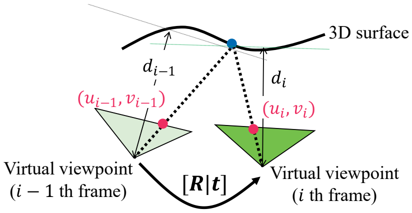

Figure 2: Relative camera motion estimation.

2.1 Offline process: 3D reconstruction

In offline processes, omnidirectional images of a target scene are captured at various locations in an area where users are supposed to move in a specific application. The omnidirectional images are utilized to reconstruct camera poses and a 3D model of the real environment using structure from motion and multi-view stereo. We employed VisualSFM [9] and CMPMVS [3] for the reconstruction. A reconstructed 3D model and camera pose information employed for our experiment are shown in Figure 1.

2.2 Online process 1: Camera motion estimation

An augmented image from the device viewpoint is displayed on the mobile device based on real-time camera pose estimation and image-based rendering.

VSLAM can be utilized for acquiring device viewpoint, where state-of-the-art methods mark high localization accuracy. However, the camera position estimated by vSLAM does occasionally not well correspond to the short-term relative motion of the camera because of the lack of robustness to texture-less scenes or behaviors during the cancellation of accumulative errors. Users recognize the difference of relative motion as the spatial (i.e. geometric) inconsistency between IAR scenes and the real world. For displaying motion parallax, it is important to estimate camera position well corresponding to the relative motion of the device rather than achieving high localization accuracy.

We reconstruct relative camera motion between two consecutive frames (see Figure 2). We directly employ device orientation $\mathbf{R}$ acquired from an inertial orientation sensor. The relative change of the camera position $\mathbf{t}$ is estimated using optical flows acquired from the images captured by the device camera. In a similar situation, Ventura et al. [7] estimated poses of upright panoramas by reducing two degrees-of-freedom (DoF); meanwhile we estimate 3-DoF relative camera translation under the constraint of given 3-DoF orientation. In our case, the position can be estimated using minimal two pairs of flows; therefore the system performs robustly in environment with less textures.

[*] e-mail: okura@am.sanken.osaka-u.ac.jp

[†] e-mail: {tomoka-s, norihi-k, yokoya}@is.naist.jp

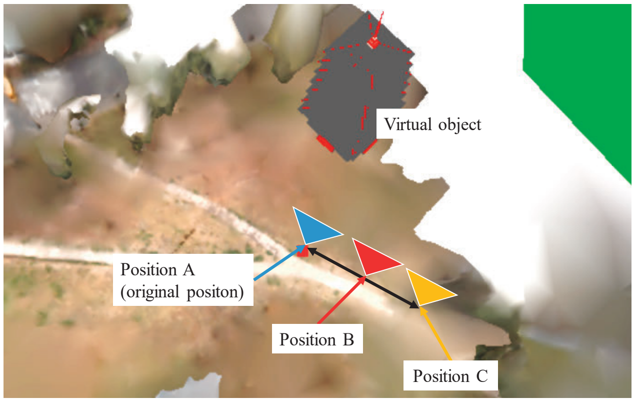

Figure 3: A top view of the experimental environment.

Let the corresponding points between the $(i-1)$-th and i-th frames be (u_{i-1}, v_{i-1}) and (u_i, v_i) assuming the intrinsic camera parameter is the identity matrix. The scene depths of these points are denoted as d_{i-1} and d_i, respectively. We assume that d_{i-1}, the scene depth in the $(i-1)$-th frame, is known based on the reconstructed 3D model and the camera pose in the $(i-1)$-th frame. The camera translation $\mathbf{t}$ as well as the depth value d_i in the current frame are estimated based on the following equation.

$$\begin{pmatrix} \frac{d_i u_i}{f} \\ \frac{d_i v_i}{f} \\ d_i \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{d_{i-1} u_{i-1}}{f} \\ \frac{d_{i-1} v_{i-1}}{f} \\ d_{i-1} \\ 1 \end{pmatrix}. \qquad (1)$$

In our implementation, $\mathbf{t}$ is estimated through a RANSAC-based outlier rejection process inputting optical flows computed by Lucas-Kanade method [4].

Notable accumulative errors in the estimated camera pose can appear by only employing a sequence of relative motion. To mitigate the accumulative errors, here we add ad-hoc implementations by assuming that users do not go far from the original position in a specific application in which virtual objects are seen from a limited area. Instead of directly employing $\mathbf{t}$ estimated using optical flows, the position of the viewpoint is modified to gradually get closer to the original position.

2.3 Online process 2: Image-based rendering with virtual object superimposition

View-dependent texture mapping (VDTM) [2], an image-based rendering technique utilizing multi-view images and a 3D model, generates a scene from the estimated position of the device camera. Onto the image generated by VDTM, virtual objects are then rendered through a traditional graphics pipeline. The relative pose of virtual objects and the reconstructed 3D models are fixed; therefore geometric misalignment between the virtual objects and the pre-captured scenes does not occur principally. Note that, to generate more photorealistic IAR scenes, static virtual objects can be superimposed offline onto pre-captured omnidirectional images [6].

3 Preliminary Experiment

We conducted a preliminary experiment to confirm whether the motion parallax is plausibly represented by the proposed system.

Experimental conditions. We captured 30 images using an omnidirectional camera, Ladybug3 (Point Grey Research, Inc.), in an area where users are supposed to move. A virtual object of an ancient tower was superimposed onto the VDTM images. The online process was performed on a mobile device, Surface Pro 3 (Microsoft Corporation; Core i7 1.7 GHz, 8GB RAM), and the resolution of IAR scenes was 640×480 pixels. We moved the device from the original position (Position A) to Position C as illustrated in Figure 3. In the initial state, the device was mounted on a jig placed at the original position and orientation.

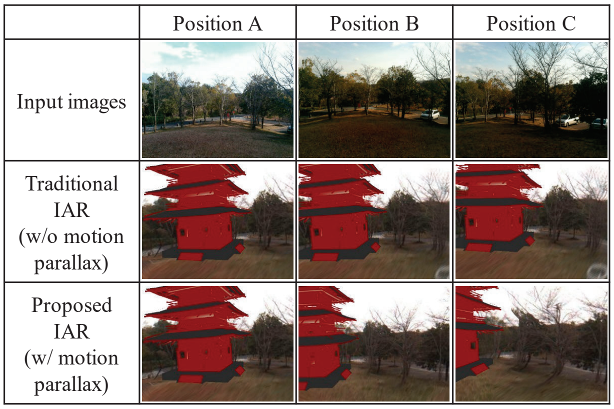
Position A
Position B
Position C
Input images
Traditional IAR (w/o motion parallax)
Proposed IAR (w/ motion parallax)

Figure 4: Results with/without motion parallax.

Experimental results. Figure 4 shows IAR scenes generated by the proposed and a traditional IAR [8], where the virtual viewpoint of traditional IAR was fixed on Position A. The traditional IAR does not represent the viewpoint change. The IAR scenes generated by the proposed system include a motion parallax effect, while the position of the virtual viewpoints mostly corresponded to the actual viewpoint (compare with input images). Our VDTM implementation performed at approximately 40 fps. Note that PTAMM [1], a vSLAM algorithm, failed to perform the localization in this environment due to the lack of informative feature points.

4 Conclusions and Future Work

Traditional IAR systems [5, 8] do not represent user's viewpoint change; therefore large spatial inconsistency can appear between IAR scenes and the real world when users change the viewpoint. The proposed system represents the viewpoint change in IAR scenes using image-based rendering, where a virtual viewpoint is acquired based on a relative motion estimated using an inertial sensor and optical flows. The experiment showed that the proposed approach presented IAR scenes including the motion parallax according to the user's behavior. We plan to evaluate the performance of the proposed approach through a subjective experiment, and actualize IAR systems enabling viewpoint change in wider areas.

Acknowledgements

This work was supported in part by MIC/SCOPE #162107002.

References

[1] R. Castle, G. Klein, and D. W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *Proc. 12th IEEE Int'l Symp. on Wearable Computers (ISWC'08)*, pages 15–22, 2008.

[2] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proc. ACM SIGGRAPH'96*, pages 11–20, 1996.

[3] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Proc. 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'11)*, pages 3121–3128, 2011.

[4] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. Int'l Joint Conf. on Artificial Intelligence (IJCAI'81)*, pages 674–679, 1981.

[5] F. Okura, T. Akaguma, T. Sato, and N. Yokoya. Addressing temporal inconsistency in indirect augmented reality. *Multimedia Tools and Applications*, Online first, DOI: 10.1007/s11042-015-3222-0, 2016.

[6] F. Okura, M. Kanbara, and N. Yokoya. Mixed-reality world exploration using image-based rendering. *ACM Journal on Computing and Cultural Heritage*, 8(2):9, 2015.

[7] J. Ventura and T. Höllerer. Structure and motion in urban environments using upright panoramas. *Virtual Reality*, 17(2):147–156, 2013.

[8] J. Wither, Y.-T. Tsai, and R. Azuma. Indirect augmented reality. *Computers & Graphics*, 35(4):810–822, 2011.

[9] C. Wu. Towards linear-time incremental structure from motion. In *Proc. 2013 Int'l Conf. on 3D Vision (3DV'13)*, pages 127–134, 2013.