摘要

相机定位是指通过分析各种传感器的数据——例如从相机中得到的视觉信息和从惯性测量单元得到的惯性信息等，来推断出相机的六自由度位姿的技术。该技术在增强现实、虚拟现实、机器人和自动驾驶等领域都发挥着极其重要的作用。

与此同时，在应用于增强现实场景的相机定位中，增强现实设备上渲染导致的延迟通常需要使用数十甚至上百毫秒前获取的传感器数据来预测头部运动，以避免虚拟内容与物理世界之间的错位，这种错位会导致用户产生时间延迟感和头晕的症状。因此，需要一种六自由度运动预测方法来补偿渲染延迟。

本文在进行了广泛的调研和全面的实验的基础上，围绕增强现实场景下如何进行准确、流畅和鲁棒的相机定位，以及如何利用六自由度运动预测方法来补偿渲染延迟这两个问题，提出了一整套技术框架，实现了具有强烈沉浸感的增强现实体验。具体而言，本文的技术贡献如下：

(1) 针对增强现实场景下的定位需求，提出了一套技术方案，通过有效结合预先完成的大规模三维重建、快速实时的局部相机定位和后台准确的全局相机定位等多项技术，实现了在现有的设备——如手机、平板和增强现实眼镜上，低成本地完成准确、流畅和鲁棒的相机定位任务，从而支持增强现实体验工程的开发。同时在该技术方案的基础上，和清华艺术博物馆与河南冶铁博物馆合作，完成了增强现实文旅导览项目的开发工作。具有优势的理论性能和良好的实地体验效果，都展示了该技术的高可用性和场景泛化性，具备较高的商业价值。

(2) 针对增强现实设备上渲染导致的延迟，提出了六自由度运动预测这个新任务，并定义了该任务的目标和对应的评估指标。提出了一种运动不确定性编解码网络（MOtion UNcerTainty encode decode network, MOUNT）来预测一段时间后的相机位姿，其能够估计输入数据中的不确定性，并预测输出位姿的不确定性，以提高预测结果的精确性和平滑性。MOUNT 具有合理的数据预处理方法和架构，可以在无监督的情况下从训练数据中学习输出位姿的不确定性预测。在公共数据集 EuRoC（European Community, EuRoC）和为该任务专门设计的数据集上分别进行的详细实验，不仅定量地证明了 MOUNT 在准确性和平滑性上显著优于传统方法，而且展示了 MOUNT 可以带来的对沉浸式增强现实效果的极大提升，相关消融实验和用户研究也支持这一结果。

关键词：增强现实；相机定位；延迟补偿；六自由度运动预测

Research on Camera Localization and Delay Compensation Methods in AR Scenes

Haoran Chen (Computer Science and Technology(Intelligent science and Technology))
Directed by: Assistant Prof. Boxin Shi

ABSTRACT

Camera localization refers to the technique of inferring the six-degree-of-freedom (6DoF) pose of the camera by analyzing data from various sensors, such as visual information obtained from the camera and inertial information obtained from the inertial measurement unit (IMU). This technology plays an extremely important role in fields such as augmented reality (AR), virtual reality (VR), robotics, and autonomous driving.

At the same time, in the field of camera localization applied to AR, the delay of rendering on AR devices requires prediction of head motion using sensor data acquired tens of even one hundred milliseconds ago to avoid misalignment between the virtual content and the physical world, where the misalignment will lead to a sense of time latency and dizziness for users. Therefore, a 6DoF motion prediction method is needed to compensate for rendering delays.

Based on extensive research and full-scale experiments, this thesis focuses on how to perform accurate, smooth, and robust camera localization in AR scenes and how to use 6DoF motion prediction methods to compensate for rendering delays. A complete set of technical frameworks is put forward to realize the AR experience with a strong sense of immersion. Specifically, the contributions of this thesis are as follows:

(1) In response to the localization requirements in augmented reality scenarios, a technical solution has been proposed that effectively combines large-scale pre-accomplished 3D reconstruction, rapid real-time local camera localization, and accurate back-end global camera localization. This solution achieves precise, smooth, and robust camera localization tasks at a low cost on existing devices such as smartphones, tablets, and augmented reality glasses, thereby supporting the development of augmented reality experience projects. Furthermore, based on this technical solution, collaborations have been established with Tsinghua Art Museum and Henan Metallurgy Museum to successfully develop augmented reality cultural tourism guide projects. The advantageous

theoretical performance and satisfactory on-site experience outcomes demonstrate the high applicability and scene generalization capabilities of this technology, indicating considerable commercial value.

(2) In light of the delay caused by rendering on augmented reality devices, a novel task called six-degrees-of-freedom (6-DoF) motion prediction has been proposed, along with the definition of its objectives and corresponding evaluation metrics. A MOtion UNcerTainty encode decode network (MOUNT) is introduced to predict camera poses after a certain period, which estimates the uncertainty in the input data and predicts the uncertainty in the output poses, thereby enhancing the precision and smoothness of the predicted results. MOUNT possesses a reasonable data preprocessing approach and architecture, allowing it to learn the uncertainty prediction of output poses from training data in an unsupervised manner. Detailed experiments were conducted on both the public EuRoC (European Robotics Challenge) dataset and a dataset specifically designed for this task. The results not only quantitatively demonstrated that MOUNT significantly outperforms conventional methods in terms of accuracy and smoothness but also showcased the considerable improvement in immersive augmented reality experiences that MOUNT can provide. Ablation experiments and user studies further corroborate these findings.

KEY WORDS: augmented reality; camera localization; delay compensation; 6DoF motion prediction