11-2023

Demo abstract: VGGlass - Demonstrating visual grounding and localization synergy with a LiDAR-enabled smart-glass

Darshana RATHNAYAKE
Singapore Management University

Dulanga WEERAKOON
Singapore Management University, mweerakoon.2019@phdcs.smu.edu.sg

Meeralakshmi RADHAKRISHNAN
University of Technology Sydney, meeralakshmi.radhakrishnan@uts.edu.au

Vigneshwaran SUBBARAJU
Institute for High Performance Computing

Inseok HWANG
Postech

See next page for additional authors

Citation

RATHNAYAKE, Darshana; WEERAKOON, Dulanga; RADHAKRISHNAN, Meeralakshmi; SUBBARAJU, Vigneshwaran; HWANG, Inseok; and MISRA, Archan. Demo abstract: VGGlass - Demonstrating visual grounding and localization synergy with a LiDAR-enabled smart-glass. (2023). *SenSys '23: Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems, Istanbul Turkiye, November 12-17*. 492-493.

Available at: https://ink.library.smu.edu.sg/sis_research/9221

Author

Darshana RATHNAYAKE, Dulanga WEERAKOON, Meeralakshmi RADHAKRISHNAN, Vigneshwaran SUBBARAJU, Inseok HWANG, and Archan MISRA

Demo Abstract: *VGGlass* - Demonstrating Visual Grounding and Localization Synergy with a LiDAR-enabled Smart-Glass

Darshana Rathnayake
Singapore Management University
Singapore
darshanakg.2021@phdcs.smu.edu.sg

Dulanga Weerakoon
Singapore Management University
Singapore
mweerakoon.2019@phdcs.smu.edu.sg

Meera Radhakrishnan
University of Technology Sydney
Australia
meeralakshmi.radhakrishnan@uts.edu.au

Vigneshwaran Subbaraju
IHPC, A*STAR, Singapore
vigneshwaran_subbaraju@ihpc.a-star.edu.sg

Inseok Hwang
Department of Computer Science and Engineering, POSTECH, South Korea
inseokh@postech.ac.kr

Archan Misra
Singapore Management University
Singapore
archanm@smu.edu.sg

ABSTRACT

This work demonstrates the *VGGlass* system, which simultaneously interprets human instructions for a target acquisition task and determines the precise 3D positions of both user and the target object. This is achieved by utilizing LiDARs mounted in the infrastructure and a smart glass device worn by the user. Key to our system is the union of LiDAR-based localization termed *LiLOC* and a multi-modal visual grounding approach termed *RealG(2)In-Lite*. To demonstrate the system, we use Intel RealSense L515 cameras and a Microsoft HoloLens 2, as the user devices. *VGGlass* is able to: a) track the user in real-time in a global coordinate system, and b) locate target objects referred by natural language and pointing gestures.

KEYWORDS

Multi-modal interaction, 3D Localization, Visual Grounding

ACM Reference Format:
Darshana Rathnayake, Dulanga Weerakoon, Meera Radhakrishnan, Vigneshwaran Subbaraju, Inseok Hwang, and Archan Misra. 2023. Demo Abstract: *VGGlass* - Demonstrating Visual Grounding and Localization Synergy with a LiDAR-enabled Smart-Glass. In *The 21st ACM Conference on Embedded Networked Sensor Systems (SenSys '23), November 12–17, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3625687.3628407

1 INTRODUCTION

Augmented reality (AR) applications have evolved to encompass various components with the motive of enhancing user experiences in real-world scenarios. Over the past decade, individual studies on these components led to substantial performance enhancements, owing to the introduction of innovative sensing mechanisms and the acceptance of multiple input and output modalities. The study delves into the interplay between human instruction comprehension and spatial localization within dynamic environments, utilizing an AR application. The demonstration integrates two prior works, namely COSM2IC [2] and LiLOC [1], to foster improved collaboration among multiple users/robots. For example, consider a scenario as shown in Figure 1. Here, the user wearing a smart glass can instruct the system to localize an object within his/her vicinity. Another user/robot, may use a localization strategy to locate the specified object's location in a global coordinate system. Building such a system requires a synergy between a precise indoor localization service as well as a system for understanding natural human instruction, which consists of both language and gestural cues.

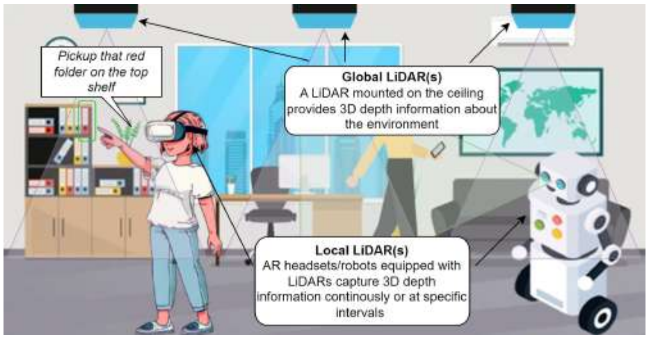
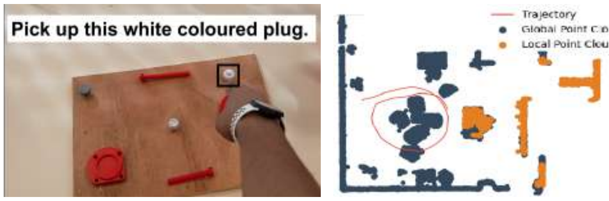

Pickup that red folder on the top shelf
Global LiDAR(s)
A LiDAR mounted on the ceiling provides 3D depth information about the environment
Local LiDAR(s)
AR headsets/robots equipped with LiDARs capture 3D depth information continously or at specific intervals

Figure 1: Motivating Scenario of an AR-enabled Human-Robot Collaborative Environment



(a) Object acquisition instruction **(b) Estimated trajectory**

Figure 2: Examples of object acquisition task comprehension and trajectory estimation

Mixed Reality headsets such as HoloLens are embedded with LiDAR sensors to accurately reconstruct the 3D environment. However, simultaneous localization of multiple users in a single coordinate system is required when multiple users/robots are collaborating in the environment. Consequently, the localization task becomes challenging as there is no unified coordinate system whereas each user reconstructs and navigates within their coordinate system. We

SenSys '23, November 12–17, 2023, Istanbul, Turkiye
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0414-7/23/11.
https://doi.org/10.1145/3625687.3628407

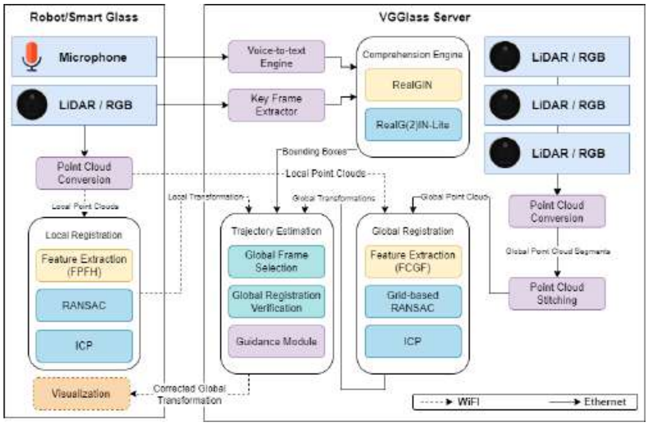


Figure 3: *VGGlass* System Architecture

use LiLOC [1] to localize the users in the environment within a single coordinate system (i.e., a global coordinate system). Instead of retaining a pre-computed map of the environment, LiLOC updates the map periodically thereby enabling localization in dynamic environments (e.g., Figure 2b). LiLOC can accurately determine the location and pose of users, achieving an error of no more than 7.4 cm and 3.2 degrees, respectively, in 84% of cases. Often in AR applications, the users also tend to interact and identify with the objects around them using natural language instructions and pointing gestures. Thus, we use a multi-modal Visual Grounding (VG) model, namely RealG(2)In-Lite proposed in COSM2IC [2]. RealG(2)In-Lite can identify a target object referred to by a natural language instruction and a pointing gesture with an accuracy of 78.9% on the COSM2IC dataset. However, RealG(2)In-Lite itself is limited to identifying the target object only from the user's point-of-view (his local coordinate system). Thus, we showcase the interplay between multi-modal VG and Global LiDAR-based localization in our *VGGlass* system, which not only tracks the user in a global coordinate system but also tracks target objects referred to by the user through language instructions and pointing gestures.

2 SYSTEM DESIGN

Figure 3 illustrates the system architecture of *VGGlass*. It consists of a set of infrastructural-mounted RealSense L515 as *LiDARs* that capture the Global Point-Cloud (GPC). The secondary device can be a smart glass, such as Microsoft HoloLens 2 or a robot with its own LiDAR (or depth) sensor. A microphone and an RGB camera are required to issue the instructions for the collaboration. Thus, the person who instructs the other person/robot will wear the HoloLense 2. As proposed in LiLOC, Local Registration (LR) is executed to determine the trajectory from the user's local coordinate system in the user's device. The poses calculated by LR are sent to the server application, where the Global Registration (GR) process takes place. This process aligns the Global Point Cloud (GPC) with the Local Point-Cloud (LPC) to ascertain the user's position within the global coordinate system. Additionally a global registration and a verification module validates the user's location, which is transmitted to the secondary device to maintain trajectory estimation.

Figure 2a shows a scenario where the user wearing smart glasses issues an object acquisition instruction for the other person/robot to comprehend. This instruction may contain both language and a pointing gesture. In such a case, *VGGlass* needs to identify both the global location coordinates of the user and the referred object. Thus, the Comprehension Engine (CE) gets activated to identify the local coordinates (As seen from the user's point-of-view) of the referred object. To comprehend this instruction, a microphone, RGB, and LiDAR camera streams captured from the HoloLens are sent over a TCP connection to the server application. The server application executes the Voice-To-Text engine to convert the audio instruction to a textual representation. We use the keyframe extractor proposed in COSM2IC [2] to identify a single RGB and a depth image, which are then used as inputs to the comprehension engine. We use two variant models for the CE; a) RealGIN; [3] a foundational VG model taking an RGB image and a textual instruction as input and retrained to our object acquisition use-case, and b) RealG(2)In-Lite model proposed in COSM2IC; a lightweight model accepting RGB, textual instruction and additionally depth image for the pointing gesture. The coordinates of the predicted bounding box (green box as depicted in Figure 1) are sent to the Guidance Module to convert them from the user's local coordinate system to the global coordinate system. Afterwards, the system can guide the other secondary device to find the corresponding object by comparing the user's path with the target object's coordinate in the global coordinate system.

3 DEMONSTRATION

We intend to arrange a live demonstration in a living space environment featuring the installation of multiple Intel RealSense L515 LiDAR sensors. In our setup, each LiDAR connects to a Raspberry Pi that collects and streams data to the central computer with the *VGGlass* system. A user wearing a Hololens 2 walks around the living space and issues pointing gesture and verbal commands to identify a target object (nuts, bolts, screws, plugs, etc) in the living space as seen from HoloLens. HoloLens streams the captured RGB, audio, and depth data into the same central computer running *VGGlass*. In real-time, the *VGGlass* outputs the 3D location of the user wearing the HoloLens along with the 3D location of the target.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation, Singapore under its NRF Investigatorship grant (NRF-NRFI05-2019-0007). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] Darshana Rathnayake, Meeralakshmi Radhakrishnan, Inseok Hwang, and Archan Misra. 2023. LILOC: Enabling Precise 3D Localization in Dynamic Indoor Environments using LiDARs. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*. 158–171.

[2] Dulanga Weerakoon , Vigneshwaran Subbaraju, Tuan Tran, and Archan Misra. 2022. COSM2IC: Optimizing real-time multi-modal instruction comprehension. 7, 4 (2022), 10697–10704.

[3] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. 2021. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

Received 26 Sep 2023; revised 26 Sep 2023; accepted 07 Oct 2023