Artificial Intelligence for Global Optimization of Blind 3D Reconstruction from Time-of-Flight Cameras

Keith Dillon[1]

[1]Univ. of Mississippi Medical Ctr.
E-mail: kdillon@umc.edu

We describe an approach to blind 3D reconstruction that combines AI-based assessment of the scene with local optimization for accurate reconstruction. This enables global optimization that does not require accurate ground truth in calibration or a good starting point for parameters. The approach can be used for initial calibration as well as continual online calibration and reconstruction.

1. Introduction

Traditional 3D reconstruction methods solved an inverse problem via optimization; given noisy 2D projections or depth maps, estimate the most probable 3D structure and sensor parameters by minimizing an error metric. These methods relied heavily on priors, good starting estimates, and regularization to stabilize the ill-posed nature of the problem. Machine learning brought a new paradigm—data-driven networks trained to map observations directly to 3D scene estimation. However, this discarded expert knowledge and required vast amounts of training data.

Several approaches seek to combine the advantages of traditional tools with powerful data driven models [1]. Recent Generative Pretrained Transformer (GPT) models promise general knowledge based on vast training datasets. Retrieval-Augmented Generation (RAG) [2] systems are used to provide specialized applications by augmenting GPT models with databases and tools. We describe such an approach here; conventional optimization is used as a subroutine guided by RAG-selected constraints and data.

2. Methods

We use the dataset from [3]. For the Visual Language Model, we utilize intensity images which can be accessed from a time-of-flight camera by summing all channels. Example images are given in Figure [1].

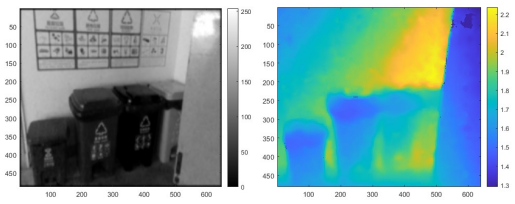

Figure 1. Example intensity image (left) and depth image (right) from time-of-flight camera [3].

The system is described in Figure [2], where the sequences of video images is treated as the database (or long term memory) for the RAG system. For the visual language model, we used the Gemini model ([4]) which supports both text and image prompts. This model used by the system as a kind of "lab assistant" to identify successive images which appear to be related, i.e., of the same scene and include similar objects, and to provide a rough estimate of the changes between the images, to be used as a starting point for optimization. A sliding window of select related images are then passed to the local optimizer for accurate reconstruction and update of intrinsic parameters.

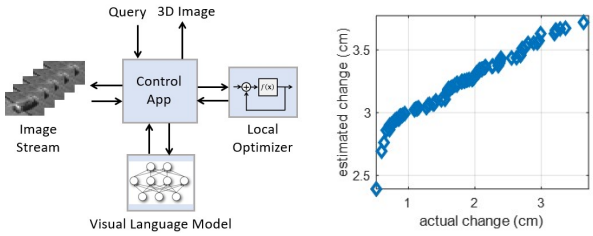

Figure 2. Diagram describing system components and interaction (left); net performance for intrinsics estimation (right).

3. Results and Conclusions

The Visual Language Model proved over 96 percent accurate at determining scene changes. Estimating initial pose estimates was more of a challenge as the model was overconfident in its ability to perform detailed calculations, despite often producing nonsensical results. Best results were achieved by casting the prompt as basic scene understanding. From that starting point, accurate optimization results could be consistently achieved. Future work includes providing the ability to directly consume and compare depth images rather than using intensity images, as well as supporting different types of cameras via new optimization tools added to the system.

References

[1] Keith Dillon, Proc. SPIE 12675, Appl. of ML (2023) 126750Z.
[2] Lewis, Patrick, et al. Adv. Neural IPS. **33** (2020): 9459-9474.
[3] Li, Yijin, et al. ECCV (2022) 619-636.
[4] https://gemini.google.com