Metadata Guided Pose Estimation for 3D Reconstruction

William Farthing[1], Compton Ross[2], Jaired Collins[3], Jing-Ru C. Cheng[3]
[1]Mathematics Department, Emory University, Atlanta, GA
[2]Department of Computer and Information Science, University of Mississippi, University, MS
[3]U.S. Army Corps of Engineers, Engineer Research and Development Center, Information Technology Lab, Vicksburg, MS

I. Introduction

Camera pose is an essential component of 3D reconstruction. Ideally, metadata is fully specified and correctly formatted, enabling direct use in downstream pipelines. In practice, this is rare: it is frequently incomplete, inconsistent, or entirely missing. Many pipelines avoid the need for qualitative evaluation by defaulting to incremental reconstruction, anchoring the first camera and estimating subsequent poses sequentially [1], [2]. This approach is computationally expensive and becomes less accurate when simplified for efficiency.

Sensor priors—from GPS and inertial navigation systems (INS)—improve both speed and accuracy, but they are difficult to obtain in practice. INS outputs vary widely in format (e.g., Euler angles, quaternions, rotation matrices), and suffer from undocumented coordinate frame conventions or axis offsets. Interpreting them correctly requires significant qualitative effort, and even then, inherent geometric ambiguity remains. Sensor error is also more common than not.

To address this problem, we introduce an interactive system that makes metadata recovery, editing, and validation a first-class component of the reconstruction workflow. The system ingests any available pose information—INS, GPS, or user input—and renders it in a 3D environment with editable intrinsics and extrinsics. When pose data is ambiguous or under-specified, the system proposes candidate configurations and filters them with computational geometry methods and lightweight Scale Invariant Feature Transform (SIFT)-based correspondence matching, all while supporting live user editing of the intrinsics and extrinsics [3].

Real-time visualization and algorithmic inference enable more accurate pose estimation in significantly less time, transforming the sensor prior obstacle into an interactive, efficient, and robust stage of the reconstruction process.

II. System Overview

A. Graphical Rendering Environment

The graphical component, developed using Qt5 and OpenGL, provides interactive visualization of camera frustums and rays. Users can project image frames to verify pose accuracy, manipulate intrinsic and extrinsic matrices via spinboxes, and perform row-column swapping and transposition operations. All matrix transformations are automatically logged, and features like view resetting support ease of use.

B. Data I/O

Our system manages data formats for raw sensor files, image data, and SIFT descriptor data. It supports direct exporting of camera poses in $K[R\,|\,t]$ format (intrinsics, rotation, and translation matrices per frame), logs of all matrix changes, and all generated point clouds. The framework is designed for downstream compatibility and applicability.

C. SIFT-based CV Pipeline

The system includes a lightweight SIFT implementation with brute-force matching for keypoints and descriptors. Using available image and metadata pairs, it reconstructs sparse point clouds on a per-frame basis. These matches and the resulting geometry are visualized directly within the 3D environment, allowing users to inspect spatial consistency and alignment. This enables approximate pose estimation based solely on image content, addressing cases where metadata is absent.

III. Software and Processing Tools

A. Preprocessing

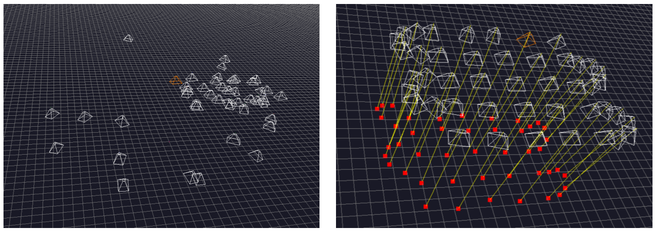
The preprocessing stage transforms raw sensor data into a unified format. We allow for multiple representations if there are metadata ambiguities GPS points are converted from latitude/longitude into Earth-centered, Earth-Fixed (ECEF) and then East-North-Up (ENU) using WGS84 ellipsoid [4]. Orientation is transformed and standardized in $\mathrm{SO}(3)$.

B. Pose Hypothesis Generation

In scenarios where sensor orientation metadata is incomplete or undetermined, the system generates exhaustive sets of plausible camera pose hypotheses. This generation process combines operations of the rotation group $\mathrm{SO}(3)$ with order permutations and chirality. For example, given undetermined Euler angles, this method generates 768 distinct orientation candidates for the sensor poses.

C. Pose Filtering and Scoring

Following hypothesis generation, the resulting pose candidates undergo geometric validation, which is a $O(n)$ process that scales well with large data sets. Poses are scored according to the distribution of their look-at direction dot products, which give angular deviations, approximate intersections, and frame consistency. Techniques such as single-point intersection checks and multi-ray surface intersection identify unlikely



(a) Synthetic Data Poses (b) Synthetic Data, Recovered

Fig. 1: Examples of incorrect camera poses (a) prior to hypothesis generation and filtering; (b) after filtering and selection (correct poses, with rays and intersections rendered).

pose candidates. Simple surface approximations (e.g. ground-plane intersection for aerial datasets) filter the remaining hypotheses so that only high-confidence pose solutions advance. For example, with standard threshold values the data in Section IV-A are filtered to eleven candidates.

IV. Results

The system has been successfully applied to a range of use cases: recovering metadata, validating sensor inputs, supporting full 3D reconstruction workflows, and rendering flight paths. We expand on three of the most emblematic examples.

A. *Metadata Recovery from Aerial Imagery*

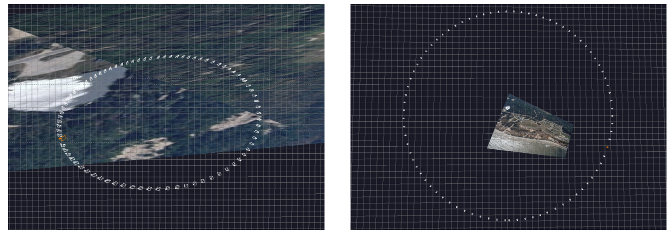
In aerial imagery with partial GPS and unknown rotation (Fig. 2), the system generated candidate poses across possible Euler angles. These were filtered using ray-intersection geometry and visual cues to discard implausible configurations. Sparse SIFT matching was then used to evaluate the remaining candidates, allowing for selection of the most consistent pose estimates. The resulting extrinsics and intrinsics were exported for use in downstream 3D reconstruction tools.

B. *Synthetic Tests*

To assess robustness, we applied the system to synthetic datasets (Fig. 1) designed to test its pose recovery capabilities. These experiments introduced controlled ambiguity and noise into the metadata, allowing us to stress test the filtering and validation pipeline under difficult conditions.

C. *Long Ground Vehicle Path*

We tested the incorporation of the software with the initial steps of a 3D reconstruction task, using image data from a long ground vehicle path. As with the aerial imagery, the poses for the camera were processed from the available metadata in dramatically less time than by hard-coding even by a domain expert. The second critical function of the metadata recovery software was evaluating sensor error in order to optimize the weighting of the pose parameters in fast bundle adjustment (Fig. 3). Lastly, the GUI allowed efficient validation of features from SuperPoint [5], matches from SuperGlue [6], and point clouds from fast bundle adjustment.



(a) Raw Aerial Input (b) Solved Aerial Data

Fig. 2: Aerial imagery: (a) with raw, inaccurate metadata as shown by the warped image projection; (b) recovered orbit with correct image projection.

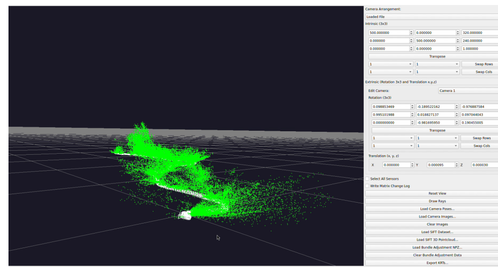


Fig. 3: Ground path poses, rendered with a rough point cloud

V. Takeaway

This work presents a system for metadata recovery in 3D reconstruction workflows. It embeds lightweight algorithms into an interactive environment for constant evaluation, efficient confirmation, and improved qualitative work. By providing a practical and robust framework for metadata recovery, this system provides a novel tool to reduce computational cost and improve accuracy in 3D reconstruction pipelines. Future work includes integrating more robust downstream SfM pipelines and applying the pose prior generation to the edge.

References

[1] J. L. Schoenberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Schoenberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html

[2] J. L. Schoenberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 501–518. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46484-8_30

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[4] National Geospatial-Intelligence Agency, "World Geodetic System 1984 (WGS 84)," National Geospatial-Intelligence Agency, Springfield, VA, Tech. Rep. NGA.STND.0036_1.0.0_WGS84, 2014, available: https://earth-info.nga.mil.

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 224–236. [Online]. Available: https://arxiv.org/abs/1712.07629

[6] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947. [Online]. Available: https://arxiv.org/abs/1911.11763