Online Manhattan Keyframe-based Dense Reconstruction from Indoor Depth Sequences

Mahdi Yazdanpour, Guoliang Fan, Weihua Sheng

School of Electrical and Computer Engineering, Oklahoma State University
{*mahdiy, guoliang.fan, weihua.sheng*}*@okstate.edu*

***Abstract*—We present an online framework for dense 3D reconstruction of indoor scenes using sequential Manhattan keyframes. We take advantage of the global geometry of indoor scenes extracted by Manhattan frames and pose optimization to enhance the accuracy and robustness of the reconstructed models. During sequential reconstruction, a Manhattan frame is extracted for each keyframe after surface normal adjustment, and used for a Manhattan keyframe-based planar alignment to initialize the surface registration while a pose graph optimization is used to refine camera poses. The final model is created by integrating the Manhattan keyframes into the unified volumetric model using refined pose estimations. Experimental results demonstrate the advantage of our geometry-based approach to reduce the cumulative registration error and overall geometric drift.**

***Index Terms*—Dense reconstruction, Manhattan keyframe (MKF), Planar alignment, Pose optimization**

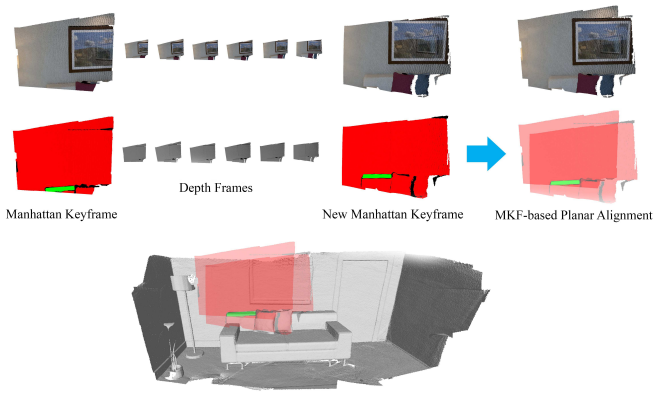

Fig. 1. The illustration of the planar alignment between two adjacent Manhattan keyframes (in red) extracted from two depth keyframes (above). A partial volumetric reconstructed model is shown below which is superimposed with the two neighboring Manhattan keyframes.

I. Introduction

Acquisition of high-fidelity 3D reconstruction of real-world scenes has become one of the most highly active research topics in robotics and computer vision. In the robotic community, the objective is often about Simultaneous Localization and Mapping and using SLAM-based frameworks to generate 3D maps with the minimum trajectory error [1], [2], [3], [4], [5]. Contrastively, in the computer vision community, researchers focus on volumetric reconstruction with the main intention of acquiring high-quality dense models [6], [7], [8], [9], [10].

With the prevalence of RGB-D sensors, there has been extensive research on 3D modeling and dense reconstruction. Many existing methods only use depth information in their reconstruction pipeline to generate 3D models. On the other hand, many methods involve both RGB and depth frames in the same pipeline to improve the accuracy and robustness of the reconstructed models. All 3D mapping and dense modeling techniques need an accurate and robust pose estimation in order to generate drift-free 3D models. This often requires visual feature matching in conjunction with bundle adjustment or pose graph optimization to minimize the reprojection error or to refine the pose estimation. These methods can perform online sequential pose optimization [7], [3], [4] or can do offline global pose optimization [9], [8]. The online approaches can produce 3D reconstructions in near real-time, but they may not have the best quality of the reconstructed models due to noise, outliers or pose tracking failures. In contrast, the offline methods can generate more accurate and detailed 3D models, however they may need more processing time and cannot be used for real-time applications. In this paper, we develop an online dense reconstruction pipeline by incorporating only depth information, which simplifies the computational flow, and by utilizing the geometric similarity between Manhattan keyframes (MKFs), which facilitates final sequential registration, as illustrated in Fig. 1. This framework reduces the cumulative registration error dramatically and generates a near drift-free 3D model without involving visual features and explicit loop closure detection. It is distinctively different from SLAM-based approaches, which involve visual odometry and use a global loop closure algorithm to correct the overall drift and generate point-based 3D maps. Our method outperforms Kintinuous [11], DVO SLAM [4], and SUN3D SfM [5]. It is also different from and outperforms Manhattan frame reconstruction (MFR) [12] which adopts Manhattan frames to create compelling volumetric models, but suffers from the accumulation of camera drift in the absence of a global pose optimization. In our most recent work, we have also proposed a local Manhattan frame growing scheme for robust reconstruction of indoor scenes [13].

II. Proposed Method

The pipeline of our approach, MFR with PGO, is shown in Fig. 2. The preprocessing step takes the raw depth stream and converts it into 3D points, followed by the estimation of

978-1-7281-3723-0/19/$31.00 ©2019 IEEE

surface normals to detect surface orientation. The next step is surface normal adjustment, which produces more persistent normal distribution. Then a Manhattan frame is estimated for each keyframe by finding three dominant orthogonal directions, and is used for a MKF-based planar alignment, which provides a reliable initialization for the surface registration. To reduce error accumulation and overall drift, we optimize the pose estimation for Manhattan keyframes. The last step is fusing the Manhattan keyframes into a volumetric representation according to the refined camera trajectory using a MKF-to-model scheme. We present four major steps in detail below.

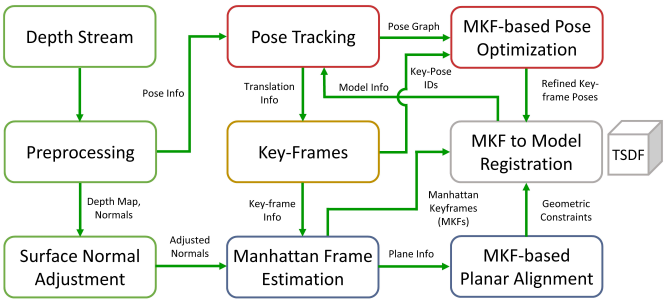

Fig. 2. Overview of our framework

A. Manhattan Frame Estimation

The representation of the structured indoor environments using Manhattan world (MW) assumption provides the reliable geometric properties for use in 3D reconstruction and scene understanding applications. In our work, we used the Manhattan frame (MF) estimation method proposed in [14] by finding the best 3D rotation matrix to transform the surface normals and align them with the principal directions in the scene. Due to noise and depth discontinuities that cause errors in normal computation, the direction of the surface normal vectors needs to be regulated. We utilize a surface normal adjustment to change the direction of the normal vectors towards the sensor location, leading to a more consistent normal distribution, which facilitates the MF estimation process. The main intention of the MF estimation is to convert normal vectors into the sparsest set of directions as follows:

$$MF = \min_{R,X} \quad \frac{1}{2}||(R \cdot N - X)||_F^2 + \lambda||X||_{1,1} \ , \qquad (1)$$

where $N \in \mathbb{R}^{3\times m}$ is the matrix of the original surface normals, $R \in SO(3)$ is the rotation matrix, and X is the sparse matrix result of applying R to matrix N. The second term in the objective function is the sum of the ℓ_1 norms of the columns in X and acts as a regularizer, and λ is the tradeoff parameter between sparsity and error sensitivity. The local minimum in this non-convex optimization is attainable via alternating optimization, where the solution for two variables R and X is updated iteratively, while the other one is kept fixed. Finally, the best estimated rotation matrix will be applied to align the surface normals to the three principal axes.

B. MKF-based Pose Graph Optimization

The odometry drift in extended scale scene reconstructions is prone to error due to the accumulation of the registration error between frames, quantization error, and noise in depth data. To alleviate error accumulation and prevent scene modeling from drifting, pose graph optimization was used to refine the pose estimations and to provide the global consistent alignments between Manhattan keyframes. For this optimization, we first construct a pose graph incrementally, where the nodes of the graph represent the estimated poses and the edges represent the relative transformation between two consecutive poses. These relative transformations between successive frames provide odometric constraints, which are used for pose optimization. In our work, we do not involve visual features nor handle loop closures. We take advantage of the geometric similarity between Manhattan keyframes to reduce the cumulative registration error. Our approach is faster than traditional SLAM-based approaches that rely on visual odometry and loop closure detection to refine the camera trajectory recursively. We optimize the pose graph through minimizing the following cost function:

$$E(T) = \underset{X}{\operatorname{argmin}} \sum_{i,j} e_{ij}^T \Omega_{ij} e_{ij}, \qquad (2)$$

$$e_{ij} = f(x_i, x_j) - d_{ij} \qquad (3)$$

where $x = \{x_0, .., x_n\}$ is a set of camera poses, $f(x_i, x_j)$ is the relative transformation between two consecutive poses calculated from camera pose tracking, d_{ij} is an observed constraint derived from two adjacent Manhattan keyframes, and Ω_{ij} represents uncertainty and is the covariance matrix of the relative transformation between consecutive frames. The main purpose of the residual cost function is to correct the pose drift by minimizing the error between the calculated measurement and observed measurement of the poses. We use `Ceres Solver` [15] to optimize this problem and minimize sequential constraints between frames.

C. MKF-based Planar Alignment

In the final stage of dense reconstruction, we propose a MKF-to-model registration based on the Manhattan keyframes. The dominant orthogonal directions of the scene are obtained based on the Manhattan world assumption and a Manhattan frame is extracted for each candidate keyframe, which is selected when a significant rotation or translation in camera poses is identified. Adopting Manhattan keyframes provides reliable geometric constraints, which helps to reduce the sequential registration error and generate an accurate dense model with minimal camera drift. A pose ID for each keyframe is stored as a retrieval index for the final recall of refined pose estimation. The pose graph optimization provides a consistent camera trajectory, which returns refined poses for all frames. For each Manhattan keyframe, the refined pose estimation is retrieved and kept for the following sequential registration. First, we perform geometric registration to efficiently align

dominant planes in two consecutive keyframes f_i^k and f_j^k. For the MKF-based planar alignment, the metric distance between the point set on two surfaces is minimized by solving:

$$T_{ij} = \underset{T_{ij}}{\operatorname{argmin}} \sum_{i,j} ||T_{ij}.(p_i) - p_j||^2, \tag{4}$$

where (P_i, P_j) is a pair of dominant planes located in two consecutive keyframes f_i^k and f_j^k, $p_i \in P_i$ and $p_j \in P_j$ are the point sets on two dominant planes, and T_{ij} is the transformation matrix that minimizes the distance between two planar surfaces. This MKF-based planar alignment enhances the speed and robustness of the final point-to-plane surface registration by providing a reliable initialization.

D. MKF-based Model Reconstruction

For the final model reconstruction, we adopt a MKF-to-model registration scheme based on the Manhattan keyframes to integrate the incoming depth keyframe to the reconstructed TSDF (Truncated Signed Distance Function) model. The TSDF is represented in GPU memory on a volumetric grid by an array of voxels. Each voxel at location p contains a signed distance TSDF value $v(\mathrm{p})$ and a voxel weight $w(\mathrm{p})$. To integrate an i^{th} incoming Manhattan keyframe into the reconstructed model, the value of each voxel is updated by:

$$v_i(\mathrm{p}) = \frac{v_{i-1}(\mathrm{p}) w_{i-1}(\mathrm{p}) + v_i(\mathrm{p}) w_i(\mathrm{p})}{w_{i-1}(\mathrm{p}) + w_i(\mathrm{p})}, \tag{5}$$

where $w_i(\mathrm{p})$ denotes the surface measurement uncertainty and is defined by

$$w_i(\mathrm{p}) = min(w_{i-1}(\mathrm{p}) + w_i(\mathrm{p}), w_{max}), \tag{6}$$

In our experiments, we set $w_i(\mathrm{p}) = 1$, as a simple average, and $w_{max} = 128$. After Manhattan keyframe integration, the final 3D model is reconstructed using the new pose estimation results retrieved from optimized camera trajectory, resulting in a volumetric surface representation of the scene.

III. Experimental Results

We evaluated our approach on the augmented ICL-NUIM dataset provided by [9], which augments the synthetic models of two indoor scenes, a living room and an office created by [16]. Our results are compared against Kintinuous [11], DVO SLAM [4], SUN3D SfM [5], and MFR [12]. To provide a quantitative comparison, the `CloudCompare` [17] is used to align our reconstructed models to the ground-truth surface provided by [16] for the living room, and to the dense point-based surface model provided by [9] for the office. We have compared our computed mean distances with errors measured by [9], as shown in Table I. The quantitative results show that our approach outperforms Kintinuous, DVO SLAM, SUN3D SfM, and MFR. In addition to the online sequential implementation, it shows that our approach significantly reduces the average mean distance of the reconstructed models (near 56% improvement over Kintinuous, 42% over DVO SLAM, 30% over SUN3D SfM, and 42% over MFR). We additionally compared our performance with robust reconstruction [9] as a reference. The models generated by our online framework have comparable quality and our method is also on par with this offline pipeline in terms of accuracy, as shown in Fig. 3.

IV. Conclusion

We presented an efficient approach for indoor scene dense reconstruction by taking advantage of the Manhattan frame estimation and pose graph optimization. The Manhattan frames are only estimated for a small set of keyframes and a MKF-based planar alignment is used to provide a reliable initialization for the final surface registration. The final model reconstruction is accomplished by using the sequentially refined camera poses by integrating the Manhattan keyframes into a unified TSDF model, leading to a volumetric representation of the scene. Experimental results show the effectiveness and robustness of our proposed approach to reduce the accumulative registration error and overall geometric drift and to generate globally consistent and reliable dense 3D models.

Acknowledgment

This work is supported in part by OCAST under grant HR18-069, NSF under grant NRI-1427345 and NIH under grant R15 AG061833.

References

[1] M. Hsiao, E. Westman, G. Zhang, and M. Kaess, "Keyframe-based dense planar SLAM," in *Proc. ICRA. IEEE*, 2017.
[2] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using kinect-style depth cameras for dense 3D modeling of indoor environments," *Int. J. Robotics Research*, vol. 31, pp. 647–663, 2012.
[3] T. Whelan, M. Kaess, H. Johannsson, J. J. L. M. Fallon, and J. McDonald, "Real-time large scale dense RGB-D SLAM with volumetric fusion," *Int. J. Robotics Research*, vol. 34, pp. 598–626, 2015.
[4] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc. IROS*, 2013.
[5] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. ICCV*, 2013.
[6] A. Dai, M. Nießner, M. Zollöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface re-integration," *ACM TOG*, vol. 36, 2017.
[7] H. Wang, J. Wang, and L. Wang, "Online reconstruction of indoor scenes from RGB-D streams," in *Proc. CVPR*, 2016.
[8] Q. Y. Zhou and V. Koltun, "Dense scene reconstruction with points of interest," *ACM TOG*, vol. 32, 2013.
[9] S. Choi, Q. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proc. CVPR*, 2015.
[10] W. Dong, Q. Wang, X. Wang, and H. Zha, "PSDF fusion: Probabilistic signed distance function for on-the-fly 3D data fusion and scene reconstruction," in *Proc. ECCV*, 2018.
[11] T. Whelan, J. McDonald, M. Kaess, M. Fallon, H. Johannsson, and J. Leonard, "Kintinuous: Spatially extended KinectFusion," in *Proc. RSS Workshop on RGB-D*, 2012.
[12] M. Yazdanpour, G. Fan, and W. Sheng, "Real-time volumetric reconstruction of Manhattan indoor scenes," in *Proc. VCIP*, 2017.
[13] M. Yazdanpour, G. Fan, and W. Sheng, "Online reconstruction of indoor scenes with Local Manhattan Frame Growing," in *Proc. CVPR Workshops*, 2019.
[14] B. Ghanem, A. Thabet, J. C. Niebles, and F. C. Heilborn, "Robust Manhattan frame estimation from a single RGB-D image," in *Proc. CVPR*, 2015.
[15] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org.
[16] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. ICRA*, 2014.
[17] D. Girardeau-Montaut. (2015) Cloudcompare: 3D point cloud and mesh processing software open source project. http://cloudcompare.org.

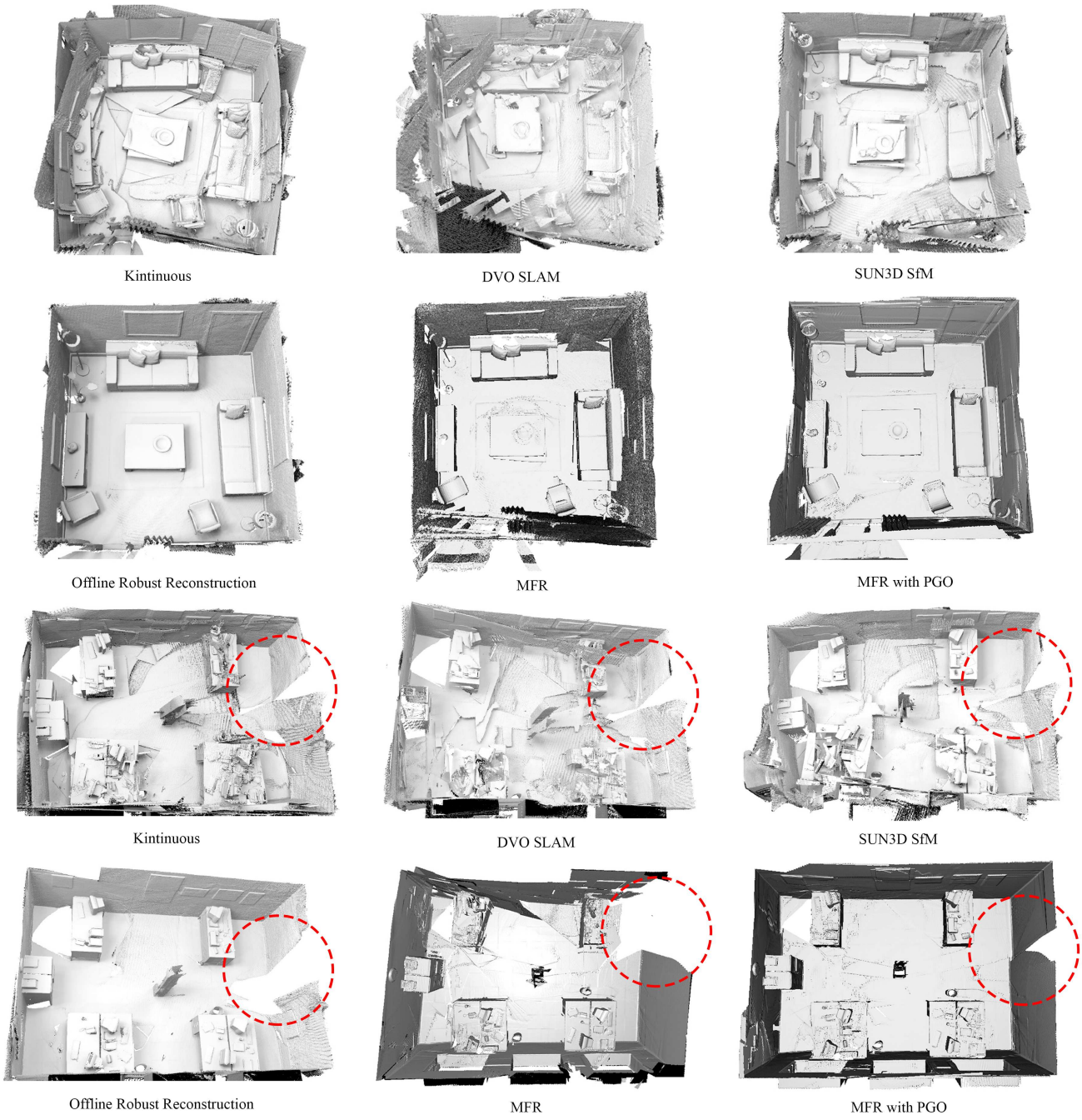

Fig. 3. Reconstructed models of Living Room1 (top) and Office1 (bottom), by Kintinuous [11], DVO SLAM [4], SUN3D SfM [5], Offline Robust Reconstruction [9], MFR [12], and our proposed MFR with PGO. Our approach has also a better performance in preserving the geometry of the planar surfaces (red circles).

TABLE I
MEAN DISTANCE OF THE RECONSTRUCTED MODELS TO THE GROUND-TRUTH SURFACE (IN METERS)

Dataset	**Kintinuous [11]**	**DVO SLAM [4]**	**SUN3D SfM [5]**	**MFR [12]**	**MFR with PGO**	**Offline Robust Reconstruction [9]**
Living Room1	0.22	0.21	0.09	0.11	0.07	0.04
Living Room2	0.14	0.06	0.07	0.09	0.07	0.07
Office1	0.13	0.11	0.13	0.12	0.04	0.03
Office2	0.13	0.10	0.09	0.17	0.06	0.04
Average	0.16	0.12	0.10	0.12	0.07	0.05