BUNDLE ICP WITH VIRTUAL DEPTH FOR HAND-HELD 3D SCANNER

Changhun Sung, Byungdeok Kim

S.LSI business, DS division, Samsung Electronics, South Korea

ABSTRACT

RGB-D cameras provide both color information and per-pixel depth information. The richness of their data present an attractive opportunity for mobile application such as Augmented Reality (AR) and 3D scanner. In this paper, we propose a general-purpose hand-held 3D scan system that combines a iterative closest point (ICP) algorithm based on a large amount of virtual information for accuracy with the advantage of a graph-based reconstruction system for robustness. First, the image graph is built with correlation between the images and the camera motion is estimated with proposed Bundle ICP method by fusing the visual information and virtual depth information. After registering the image incrementally, a sparse reconstruction model is created with global pose consistence. Finally, all the depth data from registered image are fused into a single global volume to reconstruct surface. The results of the experiment show that the proposed method has better performance than the traditional vision-based structure from motion (SFM) and tracking based reconstruction methods.

Index Terms— *3D Reconstruction , ICP, Virtual Depth*

1. INTRODUCTION

The idea of reconstructing the real world in 3D in mobile systems has gained popularity in recent years. Mixed reality such as AR and VR systems combine virtual objects with the real world. The growing digital mapping and metaverse market will demand more advanced techniques. However, it is difficult to achieve a reliable and dense 3D reconstruction from images. One popular 3D reconstruction pipeline is simultaneous localization and mapping (SLAM) [2,6,10,12].

SLAM focuses on real-time tracking and reconstruction from sequential image sets. SLAM systems are optimized for camera tracking with a spare point cloud, so they are only to produce a sparse scene reconstruction. Recently, many visual SLAM systems have come to rely on RGB-D sensors because of their ability to record rich information. One of the most notable RGB-D-based 3D reconstruction pipelines is the KinectFusion algorithm [1]. This system provides volumetric dense reconstruction of a small sized scene by combining a large amount of individual depth information into a single volumetric reconstruction. The result is that the surface model has much better quality than typical noisy raw RGB-D point clouds. Despite its good performance, KinectFusion has several limitations. One major limitation is that camera pose is estimated by frame-to-model alignment using only depth information, which can easily fail pose estimation when there are large movements between consecutive frames. Furthermore, the size of the scene to be reconstructed is restricted to a fixed small area because of limited memory size. To overcome the limitations, a common strategy is to combine RGB features into an iterative closest point (ICP)-based framework to estimate camera pose [2,7,8,9,11,19]. These feature-based methods have shown better performance than the ICP-only framework with a large shift. However, feature-based methods estimate the pose of a sensor with consecutive frames, which accumulates drift in the reconstruction model because of inaccurate feature extraction and matching. Practically, global consistency methods, such as pose graph optimization or bundle adjustment are applied to reduce the accumulated error; however, there is a limit to how much can be recovered from a bad pose estimation.

Another widely used 3D reconstruction pipeline is incremental structure to motion (SFM) [3,4,5] for reconstruction from unordered image sets. The incremental SFM process comprises a sequential step with iterative registering of new images and refining of the reconstruction using bundle adjustment (BA). Vision-based incremental SFM systems have exhibited impressive results for large-scale reconstruction with unordered internet photographs. However, vision-based SFM systems are limited in their ability to implement small scale reconstruction system such as the hand-held 3D scan system. This is because a large baseline between image pairs is required due to the accuracy of the 3D points and pose estimation.

$$\Delta z \propto \frac{z^2}{df},\ d\text{: baseline}, f\text{: focal length}, \Delta z\text{: depth error} \quad (1)$$

In this paper, we present general purpose hand-held 3D scan system that combines the advantages of a graph-based reconstruction system for robustness with a large amount of visual information based ICP for accuracy. To overcome the limitations of ICP-based methods covering narrow camera motion, the proposed method, named Bundle ICP, directly estimates the semi-global poses with incrementally registered image nodes to maximize the consistency of matched feature points across registered frames. This allows the proposed system to be more accurate under the condition of large shifts. Additionally, Bundle ICP refines the pose estimated using a virtual depth map. This enables an improvement in the accuracy of small baseline movement compared to the traditional visual SFM method.

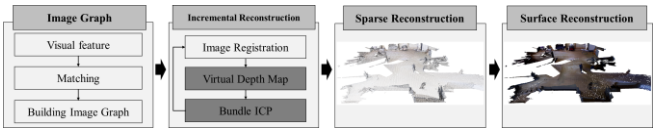

Fig. 1 Overview of Bundle ICP-based hand-held scan system

2. PROPOSED SYSTEM

The proposed 3D reconstruction system consists of three major steps: building an image graph, sparse reconstruction, and surface reconstruction. The pipeline of the proposed system is illustrated in Fig. 1. For each RGB-D frame, we first extracted the RGB feature, and then analyzed the connectivity between all possible image pairs to build an image graph. Next, we incrementally added new views using the proposed pose estimation method for the sparse reconstruction model. Finally, we estimated the surface reconstruction from a optimized sparse reconstruction model.

2.1. Image graph

The image graph plays an import role in our incremental reconstruction strategy when finding an initial image pair and selecting the next image to be added to the reconstruction model. Our proposed system consists of three steps in understanding the input image relationship. First the local feature for each image is extracted and the correspondence point between image pairs is found. Then, the image graph is built by calculating connectivity between image pairs.

Feature Extraction The image graph was constructed by finding correspondence points between two image pairs. To find the correspondence point between images obtained from hand-held sensor. The proposed system uses the most widely used feature, Scale-Invariant Feature Transform (SIFT) [13, 14]. Our system extracts visual features F_i from each image I_i in the input image set $I = \{I_i | \, i = 1,2 \dots N_I\}$

$$F_i = \{(\mathrm{x}_j, d_j) \,|\, j = 1,2 \dots N_F\} \quad (2)$$

Where x_j is the j-th local feature location $\mathrm{x}_j = (x, y)$ and d_j represents a descriptor of feature x_j.

Matching After extracting local feature from each image, the proposed system finds the correspondence point between each pair of images. Recently, graphics processing unit (GPU)-based matching techniques are widely used to perform fast feature matching. The proposed system also performs feature matching based on GPGPU [17,20]. In the matching step, Image I_i searches for feature correspondences for every feature in image I_j. The matching result between image I_i and I_j is defined as follows:

$$M_{i,j} = \{(I_i, I_j) \,|\, 0 < \, i, j < N_I \,\} \quad (3)$$

Image Graph After the matching step, our proposed system builds an image graph by calculating the connectivity between each pair of images. Each image I_i is described as node N_i in image graph and each node contains local feature location x_j and connectivity C as follows:

$$N_i = \{(\mathrm{x}_j, C) \,|\, j = 1,2 \dots N_F, i = 1,2 \dots N_I\} \quad (4)$$

The connectivity $C = \{Cor, V\}$ indicates relation information with other nodes using two properties. Cor is the sum of the correspondence point numbers with other image nodes that describe the strength of the connectivity and is used to determine the initial image pair. Visibility V is the number of 2D points that have at least one correspondence point in another image that is part of registered 3D point. V is used to determine the next image in the incremental reconstruction model. The higher the value of V is, the greater is the correlation with the registered image. Edge $E_{i,j}$ between node N_i and N_j is defined as follows:

$$E_{i,j} = \{(N_i, N_j, M_{i,j} \,|\, I_i, I_j \, \in I_{input}, M_{i,j} > \, M_{th}\} \quad (5)$$

If every number matching with the other image node is less than M_{th} , our system excludes that node from sparse reconstruction process.

Note that, in the proposed system, the initial location of incremental reconstruction is determined by the images pairs that have the strongest correlation within the input image set. In addition, the next new image is determined using the visibility score with registered images to select the highest connected image within the reconstruction model. Our system also uses an image graph to remove degraded images from input image set. Furthermore, the graph-based reconstruction system enables the placement of an additional image into image graph to expand the reconstruction model when omitting some part of the object. We demonstrate that this strategy shows better performance than tracking-based reconstruction in the Experiments section below.

2.2. Sparse Reconstruction

The process of sparse reconstruction involves fusing the individual scenes incrementally into a global model using the image graph. One important factor in the incremental reconstruction strategy is maximizing the accuracy of each pose estimation step while adding a new view. The proposed system addresses these challenging problems by combining visual and virtual depth information.

Virtual Depth Map To generate a virtual depth map (I_{vir}^{d}), our system needs to fuse all the depth data in the registered images to create a virtual depth map. This process uses the weighted average of each depth information to de-noise the depth data from multiple noisy depth sensor measurements.

$$I_{vir}^{d}(X_w) = \textstyle\sum_{I_i \in I_{reg}} w(X_w) D(X_w^i), \; X_w^i \; in \; I_i \quad (6)$$

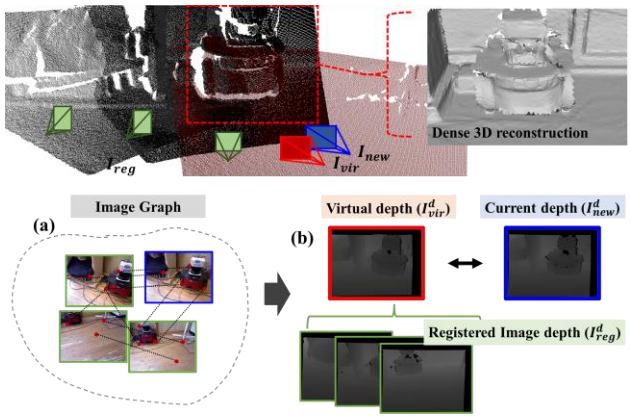

Fig. 2 Bundle ICP uses the registered images' (a) visual information and (b) virtual depth information to estimate pose

Algorithm 1: Bundle ICP algorithm

Bundle-ICP (I_{reg}, I_{new})
$M_{reg,new} = Find\ Correspondence\ Point(I_{reg}, I_{new})$
$T^*_{new} = \underset{T^w_{new}}{\operatorname{argmin}} \sum_{I_i \in I_{reg}} |T^w_i X_{i,m} - T^w_{new} X_{new,n}|^2$
$I^d_{vir} = Virtual\ Depth(T^*_{new}, I^d_{reg})$
repeat:
  $A_d = Compute\_Closest\_Points(T^*_{new}, I^d_{vir}, I^d_{new})$
  $T^*_{new} = Optimized\_Alignment(T^*_{new}, I^d_{vir}, I^d_{new})$
until(*distance_Change* $(T^*_{new} < th_{error})$ or (*Iteration* <*Max Iteration*)

where $D(X^i_w)$ represents the virtual depth value of the X^i_w, $D(X^i_w)$ is obtained by projecting 3D point X^i_w in the registered image I_i into initial current image coordinate, and $D(X^i_w) = proj(T^{new}_w X^i_w)$. Transform matrix T^{new}_w is calculated using visual-feature information. The value $w(X_w)$ is average weight at 3D point (X_w).

Bundle ICP ICP [21,22] shows a good performance when the two set points are already nearly aligned. Otherwise, the data association can converge at an incorrect local minimum. Therefore, frame-to-frame [12] or frame-to-model (F2M) ICP [1,10] methods are not suitable for dynamic application such as in hand-held scanning where dynamic sensor movement causes a small overlap between two views. By contrast, visual features can provide good correspondence points under harsh conditions. To take advantage of both types of information, the Bundle ICP combines the ICP method with visual features and virtual depth values to obtain an optimized alignment.

The input to the Bundle ICP is the image graph for the correspondence point between registered and new frames. The registered image set I_{reg} is defined as $I_{reg} = \{I_i \mid i = 1 \dots N_{reg}\}\ with\ I_i = (x_j, C)$. The output of Bundle ICP is a transform matrix between the new image and aligned sparse reconstruction model. The Bundle-ICP algorithms is described in Algorithm 1 and Fig. 2. For a new image, the Bundle ICP finds the visual correspondence points between the new Image(I_{new}) and registered image(I_{reg}) and then estimate initial new frame pose (T^*_{new}) by minimizing the least square error of the 3D point distance. After estimating the initial transformation matrix, a virtual depth image is computed with the registered image set to calculate the optimized transformation matrix using ICP algorithm.

3D point $X_{i,m}$is related to the m-th 2D point $x_{i,m}$ in image coordinate I_i. Each matching pair $(x_{i,m}, x_{new,n})$ is obtained from image edge $E_{i,new}$. T^w_i *and* T^w_{new} denote the respective transform matrices from image coordinates I_i and I_{new} to world coordinates. When a new image is added to the sparse reconstruction mode, our system also updates the 3D point set X with the depth information of the matched 2D points. Each 3D point has the list of 2D point sets and their image indices. The tracking list of 2D Point for each 3D point is used to optimize the registered image pose with global bundle adjustment. Our system also updates the visibility score of related images in image graph when registering the 3D point. To summarize, the Bundle ICP estimates the initial pose with various views of a registered image set using visual information for robustness in large shift cases and refines the initial pose with nearly aligned virtual depth map in sparse reconstruction mode using the ICP algorithm. This improves the accuracy of the pose compared to inter-frame ICP using visual features.

Global Bundle Adjustment After registering the image node with Bundle ICP, our system executes a global bundle adjustment algorithm to optimize the sparse reconstruction model. The proposed system only applies a Global Bundle Adjustment to reduce accumulated errors. Bundle ICP has an effect on the local optimization between new frame and the sparse reconstruction model. We used a popular solver, Ceres Solver [15], when implementing global bundle adjustment.

2.3. Surface Reconstruction

To generate a single global surface model, our system needs to fuse all the depth data from the registered image. The popular method in graphics is the truncated signed distance function (TSDF) [1,18] to build a mesh-based reconstruction by combining each depth scan. We denote the volume of TSDF as $V_k(X_w)$ where X_w is a 3D point in world coordinates to be reconstructed. Each voxel of the TSDF volume stores two components, namely, the truncated signed distance value $S_k(X_w)$ and weight value $W_k(X_w)$. The truncated signed distance value $S_k(X_w)$ is calculated as follows:

$$S_k(X_w) = \left(proj\left(T^i_w X_w\right) - D_i\right)\ only\ if\ \left|proj\left(T^i_w X_w\right) - D_i\right| < \mu \quad (7)$$

where T^i_w is transform matrix from world coordinates to the i-th registered image coordinate. Function $proj$ projects the 3D point into the image plane, where μ is the truncated value and D_i is depth value of I_i. Note that the TSDF-based surface reconstruction method requires maintaining the volume in the GPU memory. Aligned sparse reconstruction model-based surface reconstruction has an advantage when creating a TSDF volume. Through the result of the aligned sparse reconstruction model, the user can specify the region to be reconstructed and create the TSDF volume instead of creating all ranges of the sparse reconstruction model. This enables the reconstruction of the selected regions only, in more detail.

3. EXPERIMENTS AND RESULTS

Pose Estimation To evaluate the benefits of Bundle ICP, we evaluated the accuracy of camera motion estimation using the TUM RGB-D Benchmark Dataset [16] containing RGB-D data and ground-truth poses. We compared the proposed system in terms of the translation and rotation components. We evaluated the root mean squared translation error over all times as RPE.Tran [m], and for the rotation component we used the mean rotation error as RPE.Rot [rad].

We compared the results of Bundle ICP to four different motion estimation methods. To demonstrate the ability of the graph-based motion estimation systems, we evaluated tracking-based motion estimation methods such as F2M ICP proposed by KinectFusion. Furthermore, to show the effect of combining visual features in ICP, we evaluated F2M ICP with Visual Feature ICP (Visual ICP), which uses initial correspondence pairs of ICP from the feature matching result. To compare Bundle ICP with a state-of-the-art graph-based incremental system, we evaluated the modified COLMAP [4] system named RGB-D COLMAP, which uses 3D points from the depth information of the RGB-D sensor. Note that it is difficult for the original COLMAP to calculate an accurate 3D point in a hand-held 3D scan environment in the case of a small movement. Therefore, we used depth information instead of the triangulation step.

Table 1 summarizes the results of the four different alignment algorithms. Through experiments, F2M ICP showed higher translation and rotation errors. This was probably caused by the large inter-frame motion, as it is important for the F2M ICP methods to have sufficient inter frame overlap. Meanwhile, Visual ICP shows higher accuracy compared to the F2M ICP methods. Visual feature information helps the ICP algorithm to find the correspondence pairs between frames. However, Visual ICP has higher translation and rotation errors compared to Bundle ICP, especially using the room and plant datasets. This can be explained by the fact that Bundle ICP estimates the initial pose by considering different views of visual information in registered images.

Bundle ICP and RGB-D COLMAP showed better performance compared to tracking-based methods. The graph-based reconstruction system increased the model with highly correlated new images instead of tracking consecutive frames. Furthermore, Bundle ICP had a lower error compared to RGB-D COLMAP for all test datasets. This is because, after inferring the new frame motion with the reconstruction model, the Bundle ICP refined the camera motion of the new frame with the closest frame of the reconstruction model.

Sparse and Surface Reconstruction Fig. 3 shows the result of the sparse and surface reconstruction of our system. These results show that sparse and surface reconstruction is smooth and clean similar to the object's shape. To demonstrate our system in a practical environment, we show the result of our system surface reconstruction with datasets captured in a hand-held fashion.

Table 1: Comparison of translation and rotation error on evaluated datasets. (Tran[m], Rot[rad])

Data	F2M ICP		Visual ICP		COLMAP+RGBD		BUNDLE ICP	
	Tran	Rot	Tran	Rot	Tran	Rot	Tran	Rot
360	0.046	0.062	0.024	0.032	0.023	0.029	0.016	0.026
floor	0.036	0.024	0.020	0.018	0.017	0.021	0.013	0.012
plant	0.056	0.060	0.041	0.049	0.035	0.027	0.016	0021
room	0.055	0.039	0.045	0.051	0.027	0.020	0.015	0.017
desk	0.031	0.021	0.030	0.020	0.029	0.017	0.019	0.015

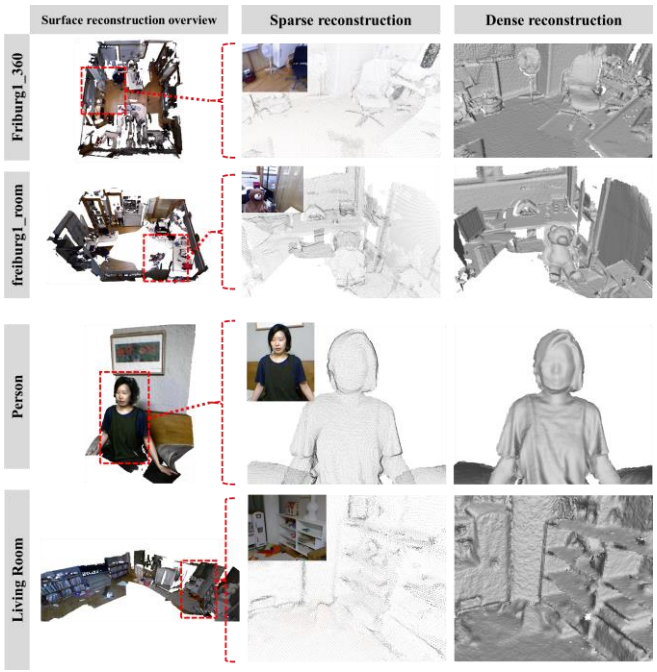

Fig. 3 Left: The result of surface reconstruction produced by proposed system; Middle: aligned sparse reconstruction; Right: surface reconstruction generated by Bundle ICP method. Small details such as the chair, bookshelf, and face in the environment are clearly reconstructed.

4. CONCLUSION

In this paper, we proposed the incremental reconstruction model-based 3D scan system to increase the accuracy and robustness of the reconstruction system under realistic conditions. Moreover, for estimating a new image pose, we presented a novel motion estimation method called Bundle ICP, which uses registered image visual information and virtual depth information. Our Bundle ICP-based 3D scan system possesses both the robustness of the incremental SFM pipeline and the accuracy gained from combining visual and virtual depth information. Extensive experimentation showed that our method produces superior results in terms of reconstruction accuracy compared to many other methods.

5.REFERENCES

[1] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," 2011 10th IEEE International Symposium on Mixed and Augmented Reality, pp. 127-136, 2011.

[2] Di K., Zhao Q., Wan W., Wang Y., Gao Y, "RGB-D SLAM Based on Extended Bundle Adjustment with 2D and 3D Information," Sensors. 2016.

[3] J. L. Schönberger and J. Frahm, "Structure-from-Motion Revisited," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113, 2016.

[4] Johannes L. Schönberger, "Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery," ETH Zürich, 2018.

[5] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 225-234, 2007.

[6] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," in IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255-1262, Oct. 2017.

[7] Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., "RGB-D mapping: using Kinect-style depth cameras for dense 3d modeling of indoor environments," Int. J. Robot. Res., 2012.

[8] Krombach, Nicola et al., "Combining Feature-Based and Direct Methods for Semi-dense Real-Time Stereo Visual Odometry." IAS, 2016.

[9] C. Wang and X. Guo, "Feature-based RGB-D camera pose optimization for real-time 3D reconstruction," Comput. Visual Media, vol. 3, no. 2, pp. 95-106, 2017.

[10] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard and J. McDonald, "Real-time large-scale dense RGB-D SLAM with volumetric fusion," Int. J. Robot. Res., vol. 34, no. 4-5, pp. 598-626, 2015.

[11] Yilmaz, O., Karakus, F., "Stereo and Kinect fusion for continuous 3D reconstruction and visual odometry," International Conference on Electronics, Computer and Computation (ICECCO), pp. 115–118, 2013.

[12] E. Mendes, P. Koch and S. Lacroix, "Icp-based pose-graph slam", Safety Security and Rescue Robotics (SSRR) 2016 IEEE International Symposium on, pp. 195-200, 2016.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," Proceedings of the Seventh IEEE International Conference on Computer Vision, pp. 1150-1157 vol.2, 1999.

[14] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors," IEEE Computer Vision and Pattern Recognition (CVPR), 2003.

[15] Agarwal, Sameer and Mierle, Keir and Others. "Ceres Solver". http://ceres-solver.org

[16] J. Sturm, N. Engelhard, F. Endres, W. Burgard and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 573-580, 2012.

[17] M. Björkman, N. Bergström and D. Kragic, "Detecting, segmenting and tracking unknown objects using multi-label MRF inference," CVIU, 118, pp. 111-127, January 2014.

[18] D. Werner, A. Al-Hamadi and P. Werner, "Truncated signed distance function: experiments on voxel size", International Conference Image Analysis and Recognition, pp. 357-364, 2014.

[19] Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., et al., "State of the Art on 3D Reconstruction with RGB-D Cameras," Computer Graphics Forum, 37(2), 625-652, 2018.

[20] Beis, J. and Lowe, D.G., "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," In Conference on Computer Vision and Pattern Recognition, Puerto Rico, pp. 1000–1006. 1997.

[21] P. Besl and N. McKay., "A method for registration of 3D shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 14(2):239-256, 1992.

[22] Sorkine O., "Least-squares rigid motion using SVD," Technical Notes, pp. 120, 2009.