Accurate Pose Estimation of a Hand-held RGBD Camera based on Sub-volume Matching for 3D Modeling

Eung-su Kim and Soon-Yong Park
School of Computer Science and Engineering, Kyungpook National University, 80, Daehak-ro, Buk-gu, Daegu, South Korea

Keywords: 3D Registration, Subvolume, Pose Estimation.

Abstract: The smoothness of the result of full body 3D reconstruction, also known as 360° reconstruction using a single hand-held sensor depends on the accuracy of the pose estimation. In this paper, we present a new idea for accurate pose estimation of such a single hand-held RGBD sensor based on subvolumetric reconstruction. In our method, we first estimate initial pose of both RGB and depth sensors through 3D coarse registration. Thereafter, in the precision matching step, we select only the keyframes for matching and estimate relative pose between them based on registration refinement. If there is a large pose estimation error between the keyframes, a subvolume is constructed using data of adjacent frames of each keyframe, and refine the final relative pose between keyframes using subvolume estimations. A series of 3D reconstruction experiments are preformed to evaluate the accuracy of the estimated pose.

1 INTRODUCTION

The rapid development of low-cost commercial sensors, such as Microsoft's Kinect, Intel realSense, and Asus Xtion Pro has resulted in broadening the research areas of modern computer vision into certain new levels. Combination of such sensors with more powerful graphics processing units (GPU) have produced many compelling results, particularly in the fields of dense 3D reconstruction, simultaneous localization and mapping (SLAM), augmented reality (AR), and structure from motion (SfM). KinectFusion (Newcombe et al., 2011) is an algorithm which permits real-time, dense volumetric 3D reconstruction of complex room-sized scenes using a hand-held Kinect depth sensor. The method uses a projection based iterative closest point (ICP) algorithm to estimate the sensor position and Truncated Signed Distance Function(TSDF) is used to facilitate the fusion of a large number of depth data used in 3D reconstruction(Nießner et al., 2013; Dai et al., 2017a). This method has a high processing speed due to low complexity of computation, but a relatively low performance pose estimation. As a solution, Dai et al. introduced an optimization method based on bundle adjustment using two-dimensional and three-dimensional data (Dai et al., 2017b). However, bundle adjustment method requires more processing time once the number of input points increases. As a solution, the authors have divided the number of frames into local areas (chunks), and performed chunk wise pose estimations. Selected keyframes that represent each area are again optimized using bundle adjustment. Maier et al. used a submap-based bundle adjustment (Maier et al., 2014). The data obtained at a similar position is determined as one submap, and the sensor pose is optimized within the submap. The combination of submaps are then globally optimized using overlapped features. However, 2D feature-based method has low accuracy in lousy illumination environment and selecting proper correspondences is ambiguous. To overcome these drawbacks, we propose a pose estimation and refinement method using keyframes and subvolumes. In our approach, we first estimate initial pose at every viewpoint using volumetric and projection-based 3D registration. Then we select an n number of keyframes among the viewpoints, and refine their pose using initial pose estimations.

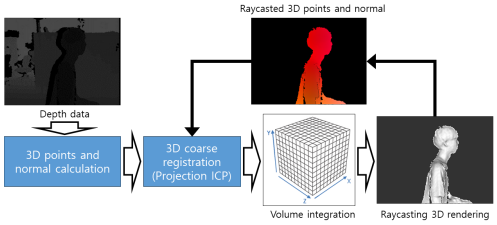

Figure 1: Flowchart of sensor initial pose estimation and 3D model reconstruction using a depth image.

Kim, E-s. and Park, S-Y.
Accurate Pose Estimation of a Hand-held RGBD Camera based on Sub-volume Matching for 3D Modeling.
DOI: 10.5220/0006857903320335
In *Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2018) - Volume 1*, pages 332-335
ISBN: 978-989-758-321-6

2 INITIAL POSE ESTIMATION USING REAL-TIME 3D REGISTRATION

2.1 3D Point and Normal Calculation

In this paper, we estimate the intial pose of a single hand-held sensor by using color and depth information at each viewpoint in a similar way to KinectFusion. The overall initial pose estimation method is schematically summarized in Figure 1. In this approach, we first reproject depth data into the 3D coordinate system to represent 2D points as 3D points. If we take D_i representing the depth of the i^{th} viewpoint, we can calculate its 3D point according to Equation (1).

$$P_i(x,y) = D_i(x,y)K^{-1}[x,y,1]^T \quad (1)$$

K represents the intrinsic parameters of the IR (infrared) camera where (x,y) represents the pixel location of the depth image. The normal of the reprojected 3D point - n_i is calculated according to Equation (2).

$$n_i(x,y) = (P_i(x+1,y) - P_i(x,y)) \times (P_i(x,y+1) - P_i(x,y)) \quad (2)$$

2.2 Sensor Pose Estimation using Projection-based Point-to-Plane ICP

ICP is a popular 3D registration algorithm that is used to minimize the difference between two point clouds. According to the traditional algorithm, the closest point of the current point cloud is calculated in its previous point cloud, and regarded as a corresponding point. A rigid transformation matrix T=$[R|t]$ is calculated that minimizes the distance between corresponding points, where R represents a 3×3 rotation matrix and t represents a 3D translation vector. Notwithstanding this traditional version achieves rigid registration with good accuracy and fast speed, it fails to register two point clouds when there are less overlapping areas and more noisy point sets

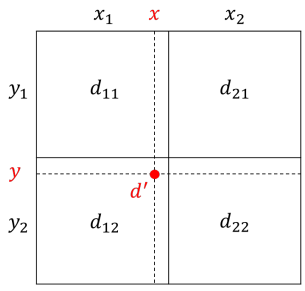

Figure 2: Calculation of arbitrary 3D point using a bilinear interpolation method.

(shape noise). As a solution, we applied a Point-to-Plane projective data association method (Chen and Medioni, 1992) to find corresponding points instead of using Euclidean distance. In general, point cloud of the i^{th} frame in the world coordinate is projected onto its $i-1^{th}$ depth image, and point of intersection is regarded as the corresponding point. However, the corresponding point is estimated by rounding due to two factors: the point of intersection is calculated as a real number, and the pixel locations of the depth image is composed of integer numbers. This increases the overall accumulating distance error and decreases the performance accuracy of the algorithm. To solve this problem, we interpolated depth data using bilinear interpolation.

As shown in Figure 2, taking already known depth values of adjacent pixels as d_{xy}, the depth value of the projected position d' can be calculated according to Equation (3).

$$d' = d_{11}\frac{(x_2-x)(y_2-y)}{(x_2-x_1)(y_2-y_1)} + d_{21}\frac{(x-x_1)(y_2-y)}{(x_2-x_1)(y_2-y_1)} + d_{12}\frac{(x_2-x)(y-y_1)}{(x_2-x_1)(y_2-y_1)} + d_{22}\frac{(x-x_1)(y-y_1)}{(x_2-x_1)(y_2-y_1)} \quad (3)$$

The converted point P'_{i-1} from $i-1^{th}$ depth data is regarded as the corresponding point of P_i. Then, transformation matrix T=$[R|t]$ is calculated by minimizing the distance between corresponding points. This transformation matrix is calculated such as Point-to-Plane ICP error metric by minimizing Equation (4),

$$T = \underset{T}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| n_i(T \cdot P_i - P'_{i-1}) \right\| \quad (4)$$

where n_i represents the normal of P_i.

2.3 Model Representation and Fusion

Depth values and the ICP-based pose information are necessary to fuse into a single consistent global coordinate. In our approach, we used the TSDF (Curless and Levoy, 1996) to fuse depth data. The volume consists of several voxels of equal size. The Signed Distance Function (SDF) of each voxel can be calculated using depth data and previously estimated pose information at each view point. These values are positive in front of the surface and negative behind the surface, with the surface dened by zero-crossing where the values change sign. We use only truncated regions around the actual surface that are referred to in this paper as TSDF. These calculated TSDF value at each viewpoint are fused according to the method proposed in (Curless and Levoy, 1996), which is summarized in Equation (5), where weight w_i is calculated according to Equation (6).

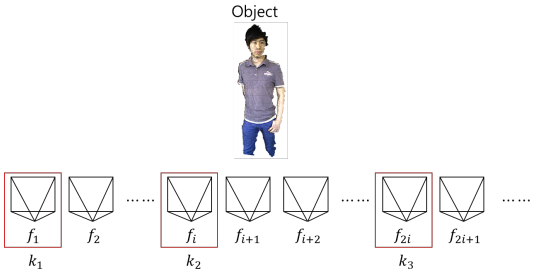

Figure 3: Selection of keyframe using uniform-sampling.

$$F^{avg} = \frac{F_{i-1}w_{i-1} + F_i w_i}{w_{i-1} + w_i} \quad (5)$$

$$w_i = min(max\ weight, w_{i-1} + 1) \quad (6)$$

2.4 Model Surface Estimation using Raycasting

The reconstructed surface at current viewpoint using accumulated TSDF value consist of significant noise. We improved the performance of pose estimation through surface information and 3D reconstruction. A simple raycaster is implemented to estimate surface information. When there is given a starting point and the direction of the ray, we traverse along the ray to extract position information of the implicit surface by observing a change in the sign of TSDF (known as zero-crossing) value. The final surface intersection point is computed applying trilinear interpolation on adjacent neighbor values of the zero-crossing. A surface normal at zero-crossing is calculated directly as the gradient of the TSDF at a zero-crossing.

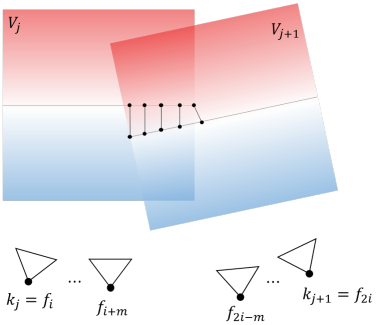

Figure 4: Keyframe pose estimation using 3D registration between subvolumes.

3 KEYFRAME-BASED 3D VOLUME REGISTRATION REFINEMENT

In the initial pose estimation step, the projected position is regarded as a corresponding point, but in precision matching, the corresponding search uses a noise robust nearest neighbor method. Moreover, 3D registration is performed only between keyframes without using all the frames. As shown in Figure 3, keyframes are selected by uniform sampling, one for each multiple frame of n, and initial pose of each selected keyframe is initialized in the world coordinate system. Then, pair-wise registration is sequentially performed in between j^{th} and $j+1^{th}$ keyframes according to Point-to-Plane ICP. However, the registration may fail due to large error in initial pose, and lack of overlapping areas. To overcome this problem, we proposed a subvolumetric-based pose estimation.

Let us consider a general situation where pose estimation between two keyframes - k_j and k_{j+1} is failed Figure 4. As shown in this figure, V_j and V_{j+1} represent two TSDF volumes generated using initial pose of first and second keyframes - k_j and k_{j+1}, respectively. Considering k_j as the world coordinate, we first selected an arbitrary m number of frames existing in between two keyframes, and sequentially matched with respect to the first keyframe (let us call these m frames as interim frames). The pose of m frames are estimated according to Point-to-Plane ICP algorithm.

Based on the estimated pose, the depth data obtained at each time interval is fused into the subvolume V_j, and the transformation relation (relative pose) between the first keyframe and respective f_{i+m} frames are stored accordingly. We repeated the same steps again for the second keyframe and its corresponding subvolume V_{j+1}, but this time selecting k_{j+1} as the world coordinate.

We reconstructed 3D surfaces at two interim frame intervals using the two subvolumes V_j and V_{j+1}, and estimated the relative pose between two subvolumes. Finally, we calculated a more refined relative pose between two keyframes using the estimated pose between the volumes.

4 RESULT

All the experiments are done in a general purpose Intel i7-7700 computer running Windows 10 (64 bit) with 16GB RAM and a Geforce GTX 980 Ti graphics card. We used an Asus XtionPro to acquire depth and RGB data. TSDF volume resolution in 3D ini-

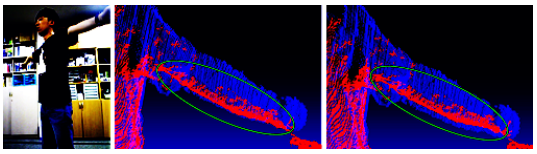

Figure 5: Experiment model (left) pose estimation result of the proposed method (middle) pose estimation result of ICP method (right).

tial pose estimation step in real-time is 512x512x512, and its area is determined experimentally between 1.2 $\sim$ 1.8m in width and height. We obtained depth and color data by moving the sensor in 360° around the model. Table.1 summarizes the nearest distance error between neighbor points for three different data sets. Figure 5 shows the experiment model we used and pose estimation results between our proposed and ICP method. We can confirm that the points are more precisely aligned in the mesh model created using our method compared to ICP method. Figure 6 shows reconstruction results and their texture mappings done using color information corresponding to each keyframe in *Data 1* data set.

Table 1: The result of Pose estimation refinement.

Data name	Frames	Nearest distance error
Data 1	773	0.73
Data 2	972	0.81
Data 3	704	0.89

5 CONCLUSIONS

In this paper, we described an accurate real time 3D pose estimation and refinement method using depth and color information of a single hand-held sensor. We first described about initial pose estimation of the sensor at real time. We selected keyframes and estimated their pose more robustly using Point-to-Plane ICP algorithm. The accuracy of the estimated pose is evaluated through experiment results such as 3D mesh modelling and texture mapping. As future work, we are planning to improve pose estimation and reconstruction results by implementing non-rigid model transformation techniques.

Figure 6: Keyframe pose estimation using 3D registration between subvolumes. Generated texture mapping model(left) and mesh model(right).

ACKNOWLEDGEMENTS

This work was supported by 'The Cross-Ministry Giga KOREA Project' grant funded by the Korea government(MSIT) (No.GK17P0300, Real-time 4D reconstruction of dynamic objects for ultra-realistic service).

REFERENCES

Chen, Y. and Medioni, G. (1992). Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155.

Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312. ACM.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017a). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1.

Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., and Theobalt, C. (2017b). Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (TOG)*, 36(3):24.

Maier, R., Sturm, J., and Cremers, D. (2014). Submap-based bundle adjustment for 3d reconstruction from rgb-d data. In *German Conference on Pattern Recognition*, pages 54–65. Springer.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. pages 127–136.

Nießner, M., Zollhöfer, M., Izadi, S., and Stamminger, M. (2013). Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):169.