Simultaneous Scene Reconstruction and Auto-calibration using Constrained Iterative Closest Point for 3D Depth Sensor Array

Meng Xi Zhu, Christian Scharfenberger, Alexander Wong, David A. Clausi
Department of Systems Design Engineering
University of Waterloo
Waterloo, Ontario N2L 3G1
Email: {mengxi.zhu, cscharfenberger, a28wong, dclausi}@uwaterloo.ca

***Abstract*—Being able to monitor a large area is essential for intelligent warehouse automation. Complete depth map of a plant floor allows Automated Guided Vehicles (AGV) to navigate the environment and safely interact with nearby people and equipment, eliminating the need for installation of guide tracks and range sensors on individual robots.**

A single camera does not have sufficient field of view or resolution to monitor a large scene, and a camera mounted on a moving platform introduces delays and blind spots that could put people at risk in busy areas. Multi-camera arrays are needed in order to reconstruct the scene from simultaneous captures. Existing iterative closest point (ICP) based algorithms fail to produce meaningful results due to ICP attempting to minimize Euclidean distance between non-matching pairs.

This paper describes a method for accurate and computationally efficient simultaneous scene reconstruction and auto-calibration using depth maps captured with multiple downward looking overhead cameras. The proposed method extends upon standard ICP algorithm by incorporating constraints imposed by the camera setup. The common field of view constraint imposed on the ICP algorithm matches a subset of points that are simultaneously in two camera's field of view. The planar constraint restricts the search space for closest points between 2 point clouds to be on a projected planar surface.

To simulate a typical warehouse environment, depth maps captured from two overhead Microsoft Kinect cameras were used to evaluate the effectiveness of the proposed algorithm. The results indicate the proposed algorithm successfully reconstructed the scene and produced auto-calibrated extrinsic camera matrix, where as standard ICP algorithm did not generate meaningful results.

I. Introduction

With autonomous robots gaining increasing popularity for use in warehouse and factory automation, various technologies are used to ensure the robots can safely navigate the environment. Visual cues or magnetic tracks are typically placed onto the floor allowing robots to traverse along a known trajectory. Changes in the warehouse / factory layout would result in costly reconfiguration of the visual cues or magnetic tracks, and dynamic environments with moving obstacles forces AGVs to be equipped with expensive sensors for obstacle avoidance. A comprehensive discussion of current use and challenges associated with AGVs are discussed by Schulze et al. [1]

With overhead depth camera arrays, depth information of the entire factory floor can be simultaneously captured, and the resulting reconstructed depth map can be used for path planning and obstacle avoidance, eliminating the need for guided tracks and distance sensors on individual robots. Such sensor array provide flexibility and robustness compared to traditional solutions. In practice, camera arrays are rarely mounted to their nominal locations, often with large positional and small angular errors. This results in a need to auto-calibrate the cameras while reconstructing a complete depth map of the whole floor.

In previous work, Microsoft Research uses projector-Kinect camera (ProCam) pairs to automatically calibrate extrinsic matrix for all units [2]. This method requires a grey-code projector to be installed with every depth camera, adding cost and complexity. Yang et al. proposed an off-line ICP auto-calibration method to align multiple Kinects under the assumption the depth cameras are looking at the same object from different angles [3]. Blais and Levine investigated using reverse calibration and very fast simulated reannealing optimzation to align point clouds of the same object [4].

The ICP algorithm proposed by Besl [5] monotonically converges to the nearest local minimum for registration and alignment, has proven to be very popular in industry due to its effectiveness and simplicity. A K-D tree implementation is proposed by Chen [6] to improve computation speed. Other variants proposed for ICP concerning different steps of the algorithm [7], [8], [9], [10] largely focuses on enhancing registration accuracy of a 3D object of interest by outlier rejection, matching techniques, and transformation estimation.

These methods all assume there is majority overlap between point clouds, which is not true in the case of depth camera arrays, where there maybe as little as 25% overlap between cameras' field of view (FOV). The standard ICP performs poorly in these situations and often converges to the wrong solution due to the algorithm's attempt to minimize the global Euclidean distance error of non-matching pairs.

A novel constrained ICP (cICP) algorithm is proposed to simultaneously align depth maps as well as calibrate the camera extrinsic matrix. Overhead camera arrays are located far apart from each other but have roughly the same downward looking camera angle, and cICP takes into consideration these characteristics to form the common FOV and planar con-

straints. cICP is shown to effectively auto-calibrate extrinsic matrices from depth map pairs and reconstructs a large scene for use in autonomous navigation and object avoidance.

The remainder of the paper is organized as follows. First an overview of the methodology is presented in Section II. Experimental results are presented and discussed in Section III. Finally conclusion and future work are discussed in Section IV.

II. Methodology

The proposed method assumes the approximate positions and orientations of each depth cameras are known beforehand. Using the approximate extrinsic parameters, captured depth maps are converted into a point cloud and the subset of points in the common FOV between 2 cameras are used in the matching step. The filtered points are projected onto a planar surface and points are matched based on minimal Euclidean distance on that surface. Finally ICP is performed to align the point clouds, and the resulting rotation matrix and translation vector are used to refine the actual pose of the cameras, as well as reconstruct the scene from different depth maps.

A. Common field of view constraint

It is crucial to enforce that only pixels in the common field of view are used for matching. This drastically reduces number of false matches, allowing cICP to reach the correct convergence. In addition, it decreases computation time by significantly reducing number of points to be matched. Therefore in the cICP a common FOV constraint is implemented as follows,

Let $v_n = \{x, y, z\}$ represent a captured depth pixel, where x,y represent the 2D location of the pixel in the depth map, and z represents the distance of the pixel to the first camera. The transformation from depth map into a point cloud captured by camera i $P_i = \{p_1, ..., p_n\}$ is

$$p_n = \begin{bmatrix} fx & 0 & cx \\ 0 & fy & cy \\ 0 & 0 & 1 \end{bmatrix}^{-1} v_n$$

where fx, fy, cx, cy are the x,y focal lengths in pixels and principal points of the camera.

Using user provided nominal camera pose as initial parameters, P_i is transformed into its respective world coordinate representation W_i via the transformation

$$w_n = \begin{bmatrix} & R_i & & t_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_n \\ 1 \end{bmatrix}$$

where R_i and t_i are the rotation matrix and translation vector of camera i with respect to a world coordinate frame.

Once the depth maps from different cameras are transformed to point clouds in a common world coordinate frame, the point cloud from camera i is inverted into the coordinate frame of another camera j with rotation matrix R_j and translation vector t_j, undergoing the following transformation:

$$p_n^{'} = \begin{bmatrix} & R_j & & t_j \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} w_n \\ 1 \end{bmatrix}$$

The points are then mapped onto a 2D coordinate frame based on the intrinsic data of camera j,

$$v_n^{'} = \begin{bmatrix} fx & 0 & cx \\ 0 & fy & cy \\ 0 & 0 & 1 \end{bmatrix} p_n^{'}$$

$p_n \in P_i$ belongs to a subset $P_i^{'}$ if its corresponding $v_n^{'}$ is within the pixel range of camera j and is the closest point to camera j for its location in the depth map.

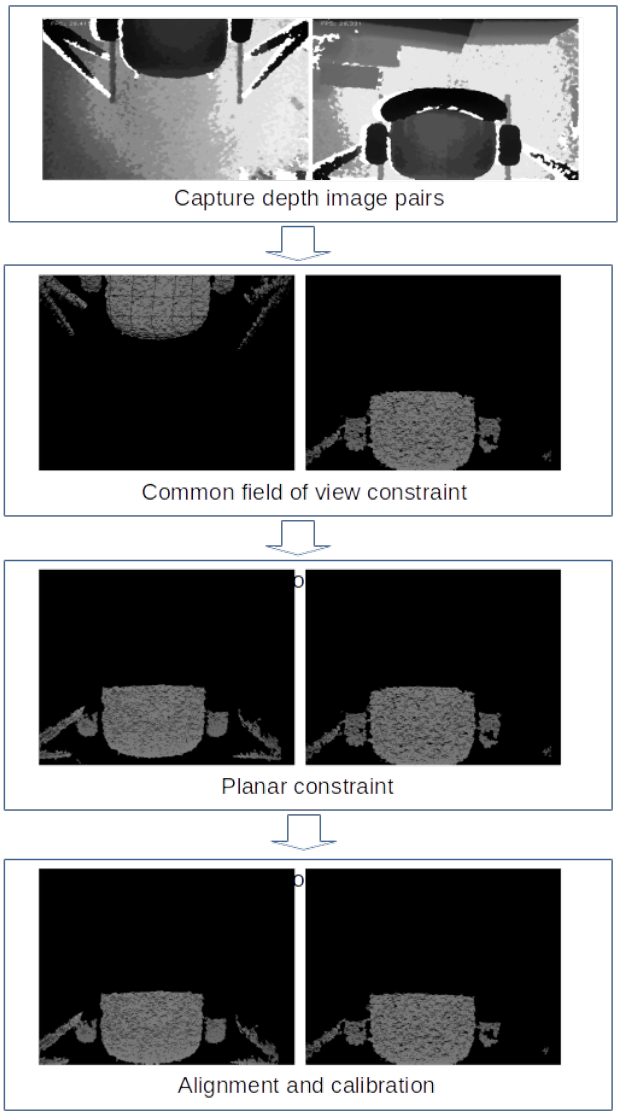


Fig. 1: Overview of constrained ICP process.

B. Planar constraint

During installation gravity aids in the alignment of orientation of depth cameras. This allows depth camera arrays installed on factory ceilings to have sufficiently similar camera angles. Using this property, P_i and P_j are projected onto a 2D plane perpendicular to the average of the camera angle, which for overhead camera installations is directly perpendicular to the floor, thus the projected points are the (x,y) values of the original 3D point if the z-axis is perpendicular to the floor.

C. cICP

Let a_n represent a point in subset $P_i^{'}$ and b_n represent a point in subset $P_j^{'}$, a_n is considered the closest neighbour to b_n if a_n and b_n's positions on the projected plane is the closest to each other. In practical implementation, a grid-based matching scheme is used to match closest points rather than a 2D kd-tree implementation since it is much faster computationally and offers similar performance.

Let a_n and b_n be the matched pair found from previous step, and m the total number of pairs found, the ICP algorithm attempt to minimize the objective function

$$[R_o, t_o] = argmin \sum_{n=1}^{m} ||R_o a_n - t_o - b_n||^2$$

where R_o and t_o are the optimal rotation matrix and translation vector that minimizes the Euclidean distance between matched pairs.

The optimal rotation matrix is found by singular value decomposition (SVD) of matrix N, where

$$N = \sum_{i=1}^{n} (a_n - \overline{a})(b_n - \overline{b})^T$$

and $\overline{a}, \overline{b}$ are the centroid of point cloud P_i and P_j respectively.

Taking the SVD of N,

$$N = U \Sigma V^T$$

where U and V are orthogonal matrices and Σ is a diagonal matrix, the optimal rotation matrix and optimal translation vector should be

$$R_o = VU^T, t_o = R_o \overline{a} - \overline{b}$$

Once optimal rotation matrix R_o is found, the rotation matrix of R_i and t_i is updated with

$$R_{i,new} = R_o R_i, t_{i,new} = t_i + t_o$$

The procedure is repeated iteratively until the change in translation or rotation matrix is below a set threshold.

III. Experimental Results

A. Experimental Setup

To quantitatively and qualitatively evaluate the proposed method, we performed a series of experiments using set of depth maps that simulate warehouse environments. A rigid mount is manufactured to hold 2 Kinect cameras parallel to each other, 63.5cm apart, and 155cm above ground. The Kinect cameras are disassembled in order to access mounting holes on the camera itself, which ensures the cameras are aligned accurately. Figure 2 shows a rendering of the mounting structure, and Figure 3 shows the mount placement of the Kinect cameras.

The depth images are captured simultaneously by both cameras. Test set 1 shown in Figure 4 has small amount of boxes in overlap region. The overlapping region of test set 2 shown in Figure 5 has more features and contains many boxes in overlap region. Test set 3 features a chair which takes up the majority of the common field of view shown in Figure 6.

Fig. 2: Rendering of Multi-Camera Mount used in Experiment

(a) View of Kinect Mount Set Up (b) Close of up Kinect Mount

Fig. 3: Kinect Mount for Proposed Experiment

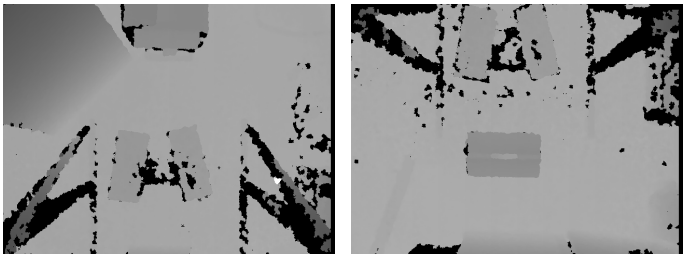


Fig. 4: Depth Image Pairs for Test Set 1

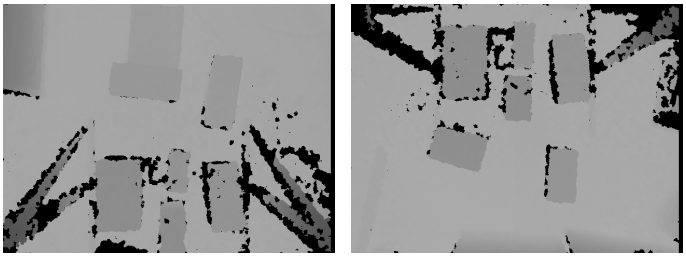


Fig. 5: Depth Image Pairs for Test Set 2

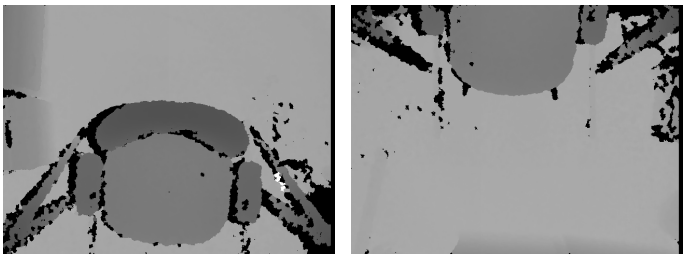


Fig. 6: Depth Image Pairs for Test Set 3

The ground truth extrinsic parameters for camera 1 is

$$R_1 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, t_1 = \begin{bmatrix} 0 \\ 0 \\ 1550 \end{bmatrix}$$

and camera 2

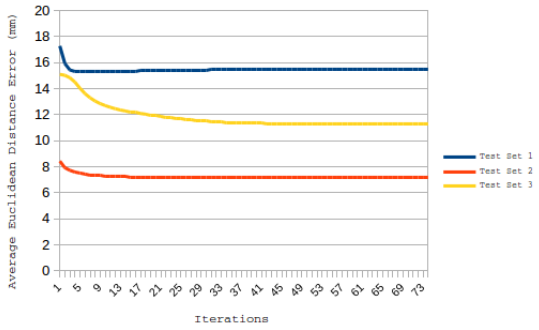
$$R_2 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, t_2 = \begin{bmatrix} 0 \\ 635 \\ 1550 \end{bmatrix}$$

where the positive Z axis points upward from the ground and unit is in mm.

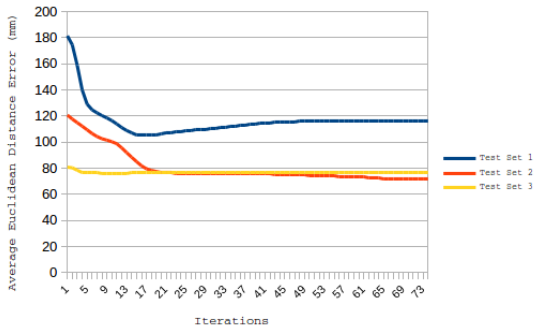
B. Results

Camera 1 is used as point of reference and origin, thus the initial parameter given for camera 1 is the same as ground truth, and only auto-calibration of camera 2 is done with respect to camera 1. Camera 2 is given an initial parameter of $t = [t_x \quad t_y \quad t_z]$ and $r = [r_x \quad r_y \quad r_z \quad \rho]$ where t is the initial translation matrix, and r the initial axis angle rotation representation of the rotation matrix and ρ is in degrees.

The extrinsic matrix for camera 2 is initialized with $t = [0 \quad 535 \quad 1550]$ and $r = [0 \quad 1 \quad 0 \quad 180]$ which represents 100mm of error between estimated (535mm) and actual (635mm) camera position, which is typical in real life scenarios.



(a) Mean Euclidean Error for cICP



(b) Mean Euclidean Error for ICP

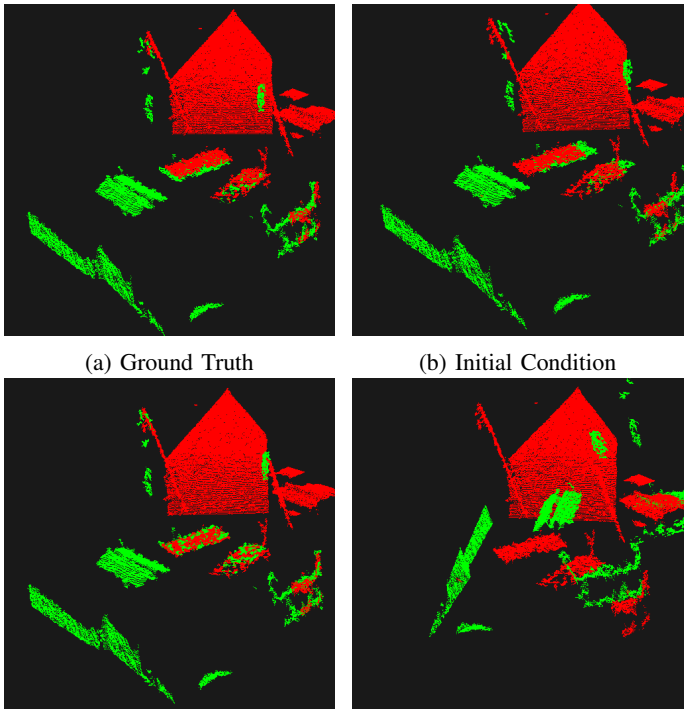
Fig. 7: Mean Euclidean Distance Error

Figure 7 shows convergence results for the 3 test sets using cICP and ICP for 75 iterations. Due to constraining the ICP algorithm to only match points in the common field of view, cICP has significantly lower average Euclidean error, as well as converge to the correct results faster for all test sets.

Figure 8, 9, 10 shows reconstructed scene using cICP and ICP algorithm. Red points are point clouds generated from camera 1, and green points are generated from camera 2. It is clear that ICP produced erroneous results while the proposed method converged to the correct solution. The deviation from ground truth may be explained by interference between Kinect cameras causing noise in the overlapping region, error introduced in the cICP algorithm due to projection onto a 2D plane, mechanical assembly error, as well as use of factory default camera intrinsic matrix.

cICP with grid matching takes approximately 100ms to run on an Intel Core 2 Duo with no hardware acceleration using pair of 640x480 depth maps, which is a significant improvement over kd-tree algorithms that takes on average 4 seconds align the same point clouds. The significant improvement in speed can be attributed to the fact that less points are used in matching as well as grid matching the points on a 2D plane. The near real time run time allows the algorithm to be run continuously, which maybe helpful in correcting disturbances to camera pose during system operation.

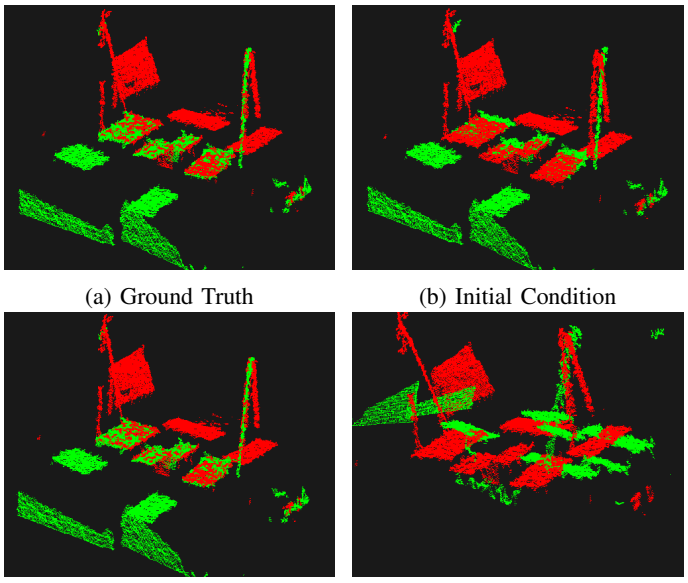
The extrinsic (homography) R_1, R_2, R_3 matrix generated



(a) Ground Truth (b) Initial Condition

(c) cICP (d) ICP

Fig. 8: Test Set 1 Scene Reconstruction using cICP and ICP. Red points are from camera 1, green points are from camera 2



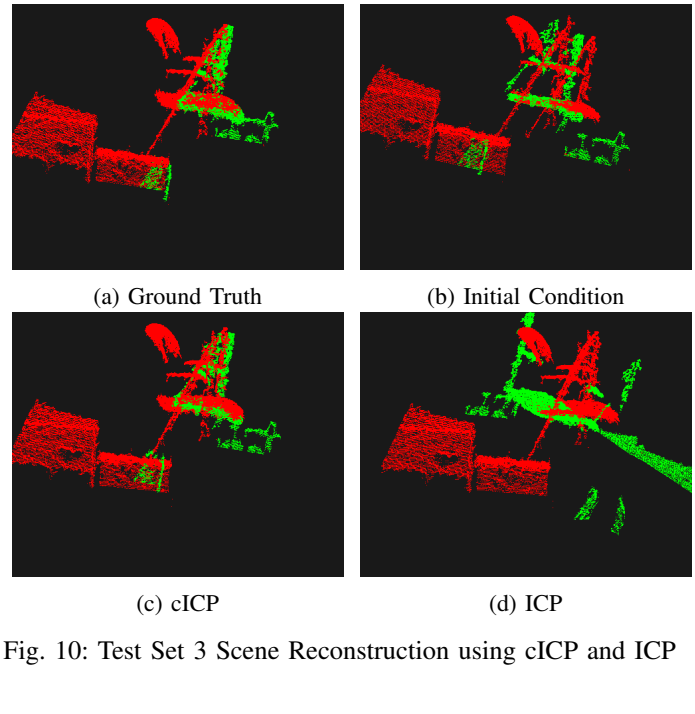
(a) Ground Truth (b) Initial Condition

(c) cICP (d) ICP

Fig. 9: Test Set 2 Scene Reconstruction using cICP and ICP.

by cICP for test set 1, 2, 3 are

$$R_1 = \begin{bmatrix} -0.9997 & -0.0069 & 0.0195 & 6.10703 \\ 0.006 & 0.9999 & -0.0004 & 640.717 \\ 0.0195 & -0.0006 & -0.9998 & 1563.99 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



(a) Ground Truth (b) Initial Condition

(c) cICP (d) ICP

Fig. 10: Test Set 3 Scene Reconstruction using cICP and ICP

$$R_2 = \begin{bmatrix} -0.9996 & -0.0264 & 0.0017 & -13.0115 \\ -0.0264 & 0.9996 & -0.0045 & 629.569 \\ -0.0016 & -0.0045 & -0.9999 & 1557.68 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_3 = \begin{bmatrix} -0.9981 & 0.0302 & 0.0518 & -16.7696 \\ 0.0283 & 0.9989 & -0.0367 & 641.31 \\ -0.0529 & -0.0352 & -0.9979 & 1536.98 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Table I shows position and orientation error using auto-calibrated camera pose from cICP and ICP. Position error is the Euclidean distance between calculated camera location compared to ground truth camera location, and orientation error is the difference between calculated camera look-at vector compared to ground truth camera look-at vector.

TABLE I: Auto-calibration Results using cICP and ICP

Image Set	cICP		ICP	
	$\Delta position$	$\Delta orientation$	$\Delta position$	$\Delta orientation$
1	16.3mm	0^o	551.6mm	154.89^o
2	16.1mm	1.14^o	383.69mm	134.29^o
3	22.1mm	1.40^o	68.86mm	158.10^o

The auto-calibrated homography using cICP is sufficiently close to ground truth, and the deviation from ground truth may be explained by interference between Kinect cameras causing noise in the overlapping region, error introduced in the cICP algorithm due to projection onto a 2D plane, mechanical assembly error, as well as use of factory default camera intrinsic matrix.

Table II show different initialization parameters and final convergence results. ICP in all cases did not generate meaningful results.

It is shown that the proposed method returned reasonable extrinsic matrices for different image sets and initial parameters. It is also shown that the proposed method is sensitive to

TABLE II: Results of cICP using Different Initialization Parameters

Image Set	Initial Conditions				Errors	
	t_x	t_y	t_z	ρ	Δposition (mm)	Δorientation
1	0	635	1550	180	14.0	1.2^o
2	0	635	1550	180	14.4	0.4^o
3	0	635	1550	180	16.4	0.5^o
1	0	600	1550	180	15.4	1.1^o
2	0	600	1550	180	14.9	0.4^o
3	0	600	1550	180	21.0	0.6^o
1	0	550	1550	180	15.9	1.1^o
2	0	550	1550	180	17.5	0.6^o
3	0	550	1550	180	21.1	0.5^o
1	0	635	1550	177.1	59.5	1.3^o
2	0	635	1550	177.1	26.0	0.4^o
3	0	635	1550	177.1	28.4	0.3^o
1	0	635	1550	174.3	84.7	3.8^o
2	0	635	1550	174.3	37.5	0.4^o
3	0	635	1550	174.3	22.2	0.2^o
1	0	635	1550	168.5	60.3	10.0^o
2	0	635	1550	168.5	223.6	7.3^o
3	0	635	1550	168.5	25.4	0.1^o

initial camera orientation, especially for depth map pairs that contain small amount of features in the overlap region.

It is also empirically determined that the amount of translation error that can be tolerated depends on the objects in the overlapping region. In general the more distinct an object's height profile and smaller the size, the closer the initial translation parameters needs to be to the ground truth for convergence.

IV. Conclusion

cICP is able to reconstruct a scene and automatically calibrate the extrinsic matrix of 2 overhead depth camera array. cICP outperforms standard ICP algorithm which generates no real solutions. The method constrains points used in the matching step to be in the overlapping field of view, which reduces amount of non-matching pairs and increases calibration accuracy. Furthermore the planar constraint matches closest points based on their projected coordinates on a 2D plane, which significantly reduces computation time and allow the algorithm to run in real-time. The proposed method may be useful in applications such as scene reconstruction of large areas for robotics and automation.

Future works involve investigating into the effectiveness of cICP in multi-view scene reconstruction with more than 2 overhead depth cameras at different positions. Inspired by [11], [12] cICP maybe further optimized by considering optimizing the extrinsic matrices of all depth cameras to achieve a global minimum Euclidean error between all depth map pairs.

In addition, a possible change to the cICP is to further refine the common points based on surface normal of a point as introduced by Maier-Hein et al. [13]. If a point from camera i has surface normal roughly perpendicular to the viewing direction of the camera j, it is rejected even though it may be in common field of view of both camera i and j. This is due to the fact depth cameras does not reliably detect distances to surfaces that are roughly parallel to the viewing direction.

Additionally, since the Kinect camera provides both colour and depth information, colour can be used as an additional constraint during matching to provide more robust and accurate reconstruction.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs Program, and Ontario Ministry Of Research And Innovation.

REFERENCES

[1] L. Schulze and A. Wullner, "The approach of automated guided vehicle systems," in *Service Operations and Logistics, and Informatics, 2006. SOLI '06. IEEE International Conference on*, June 2006, pp. 522–527.

[2] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira, "Roomalive: Magical experiences enabled by scalable, adaptive projector-camera units," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '14. New York, NY, USA: ACM, 2014, pp. 637–644. [Online]. Available: http://doi.acm.org/10.1145/2642918.2647383

[3] R. Yang, Y. H. Chan, R. Gong, M. Nguyen, A. Strozzi, P. Delmas, G. Gimel'farb, and R. Ababou, "Multi-kinect scene reconstruction: Calibration and depth inconsistencies," in *Image and Vision Computing New Zealand (IVCNZ), 2013 28th International Conference of*, Nov 2013, pp. 47–52.

[4] G. Blais and M. Levine, "Registering multiview range data to create 3d computer objects," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 820–824, Aug 1995.

[5] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[6] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," in *Robotics and Automation, 1991. Proceedings., 1991 IEEE International Conference on*, Apr 1991, pp. 2724–2729 vol.3.

[7] T. Masuda and N. Yokoya, "A robust method for registration and segmentation of multiple range images," *Computer Vision and Image Understanding*, vol. 61, no. 3, pp. 295 – 307, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314285710247

[8] K. Pulli, "Multiview registration for large data sets," in *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference on*, 1999, pp. 160–168.

[9] G. Sharp, S. Lee, and D. Wehe, "Icp registration using invariant features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 90–102, Jan 2002.

[10] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *Int. J. Comput. Vision*, vol. 13, no. 2, pp. 119–152, Oct 1994.

[11] R. Benjemaa and F. Schmitt, "Fast global registration of 3d sampled surfaces using a multi-z-buffer technique," in *3-D Digital Imaging and Modeling, 1997. Proceedings., International Conference on Recent Advances in*, May 1997, pp. 113–120.

[12] R. Bergevin, M. Soucy, H. Gagnon, and D. Laurendeau, "Towards a general multi-view registration technique," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 5, pp. 540–547, May 1996.

[13] L. Maier-Hein, A. Franz, T. dos Santos, M. Schmidt, M. Fangerau, H. Meinzer, and J. Fitzpatrick, "Convergent iterative closest-point algorithm to accomodate anisotropic and inhomogenous localization error," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 8, pp. 1520–1532, Aug 2012.