Multi-Kinect Skeleton Fusion for Physical Rehabilitation Monitoring

Saiyi Li[1], Pubudu N. Pathirana[2], *Senior Member, IEEE*, and Terry Caelli[3], *Fellow, IEEE*

***Abstract*— Kinect has been increasingly applied in rehabilitation as a motion capture device. However, the inherent limitations significantly hinder its further development in this important area. Although a number of Kinect fusion approaches have been proposed, only a few of them was actually considered for rehabilitation. In this paper, we propose to fuse information from multiple Kinects to achieve this. Given the specific scenario of users suffering from limited range of movements, we propose to calibrate depth cameras in multiple Kinects with 3D positions of joints on a human body rather than in a checkerboard pattern, so that patients are able to calibrate Kinects without extra support. Kalman filter is applied for skeleton-wise Kinect fusion since skeleton data (3D positions of joints) and its derivatives are preferred by physiotherapists to evaluate the exercise performance of patients. Various preliminary experiments were conducted to illustrate the accuracy of proposed calibration and fusion approach by comparing with a commercial Vicon system®, confirming the practical use of the system in rehabilitation exercise monitoring.**

I. INTRODUCTION

In physical rehabilitation, motion capture devices have been utilised to monitor and record rehabilitation exercises routines. Some of them, such as commercial Vicon system [1] with high resolution and accuracy, are too expensive and complicated to be operated and deployed by patients. In comparison, Microsoft Kinect® is a cheaper alternative for therapists and patients to monitor rehabilitation exercises on a daily basis in clinics and at home environments.

However, Kinect has some limitations. Firstly, the measurement range of one Kinect is limited to cover complete human movements. According to [2], a Kinect provides a limited field of view (FOV) of 43° in vertical and 57° in horizontal, while the effective range in z axis is 1 to 3 meters[3]. Although this range is enough for exercises of upper extremity since they usually involve minimal movement of whole body position, it may not fulfil requirements of lower body exercises [4] that involve walking. Secondly, Kinect is unable to detect occlusive joints [5], resulting in inaccurate measurements of 3D positions of these joints. Lastly, although the accuracy of a Kinect is good enough for gaming, it should be improved to measure rehabilitation exercises more accurately.

To overcome the limitations mentioned above, we propose to fuse skeleton data captured from a number of Kinects, involving two major steps, namely inter-Kinect depth camera calibration and skeleton data fusion. In the past few years, the former has been studied by many researchers. For example, Kinects were calibrated in [6] with global (all-to-all) external calibration method with a hand-made calibration bar. Similarly, a calibration tool with a pole and two boards was used [7] to calibrate three non-overlapping Kinects where plane, instead of corresponding points, was used as a common feature. In [8], a best-fit plane calibration method was applied to compute external parameters with a checkerboard. A similar approach (singular value decomposition) was used in [9] as we propose in this paper, but a checkerboard was also required. Moreover, multi-Kinect fusion has also been explored. For instance, in [10], a point cloud library was introduced for Kinect point-cloud-wise fusion to construct 3D models. Meshes captured from multiple Kinects were fused in [6] for 3D reconstruction of moving people in real-time with CUDA technology after calibrating these Kinects with a checkerboard and a global external calibration approach. A roulette wheel selection scheme was implemented [11] to select the best joint position from candidate joint positions captured from multiple Kinects.

From the above, it can be seen that the majority of Kinect calibration approaches are point-cloud-wise and require a checkerboard. However, in rehabilitation, especially telerehabilitation, it will be easier if patients are able to calibrate Kinects with joints on their bodies rather than a checkerboard due to their physical condition which limits their movement to varying degrees. Some of them may be unable to set up a checkerboard. Also, skeleton data is favoured in some rehabilitation scenarios since all necessary data, such as velocity and angle [12], [13], can be computed easily. Therefore, rather than fusing point cloud and then derive positions of joints from it, we propose to fuse skeleton data directly.

The contributions of this paper are two folds. Firstly, since skeleton data is quite noisy, we extend Umeyama et al.'s algorithm [14] so that the pose parameters, namely rotation matrices and translation vectors, between multiple Kinects can be optimised with joint positions of patients, instead of using checkerboards. Secondly, Kalman filter is implemented to use skeleton data as input and output of the fusion process to improve measurement accuracy.

The rest of the paper is organised as follow. The extension of Umeyama et al.'s algorithm and the model of the system for Kalmen filter are introduced in Section 2 and 3. Experiments and results are presented and discussed in Section 4, followed by the conclusion.

[1]Saiyi Li is with the Department of Electrical Engineering, Deakin Univerisity saiyi@deakin.edu.au

[2]Pubudu N. Pathirana is with the Department of Electrical Engineering, Deakin University pubudu.pathirana@deakin.edu.au

[3]Terry Caelli terry.caelli@deakin.edu.au

This project was funded by National ICT Australia and Deakin University

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

II. Multi-Kinect Calibration

To calibrate multi-Kinect system with $N+1$ Kinects (one reference and N reliant Kinects), the 3D positions of $M \leq 20$ joints will be utilised. The local Cartesian coordinate system of the reference Kinect is selected as the global coordinate system. The purpose of multi-Kinect calibration is to compute the optimised pose parameters between each reliant Kinect and the reference Kinect. To estimate poses of Kinect more efficiently, we propose to organise the data in matrix forms as follows.

The measurements of M joints captured in total $F \in \mathbb{Z}^+$ frames with time interval δt from the reference Kinect is organised as $P = [P_1, P_2, \ldots, P_M]^\top \in \mathbb{R}^{3MN \times F}$, where

$$P_m = \begin{bmatrix} (P_1^m)^1 & (P_1^m)^2 & \cdots & (P_1^m)^F \\ (P_2^m)^1 & (P_2^m)^2 & \cdots & (P_2^m)^F \\ \vdots & \vdots & \ddots & \vdots \\ (P_N^m)^1 & (P_N^m)^2 & \cdots & (P_N^m)^F \end{bmatrix} \in \mathbb{R}^{3N \times F}, \quad (1)$$

and $(P_n^m)^f = [x, y, z]^\top$ with $m = 1, 2, \ldots, M$, $n = 1, 2, \ldots, N$ and $f = 1, 2, \ldots, F$. Here $\top$ is a transposition and x, y and z are coordinates along the X, Y and Z axis of a Cartesian coordinate system respectively. In addition, $(P_1^m)^f = (P_2^m)^f = \cdots = (P_N^m)^f$.

At the same time, the measurements of these joints from reliant Kinects are notated as $Q = [Q_1, Q_2, \ldots, Q_M]^\top \in \mathbb{R}^{3MN \times F}$, which shares the same structure as P, but $(Q_1^m)^f, (Q_2^m)^f, \cdots, (Q_N^m)^f$ are not necessarily the same.

As for rotation matrices, we organised them as $R = diag\{R_1, R_2, \ldots, R_N\} \in \mathbb{R}^{3N \times 3N}$ ($diag\{\cdot\}$ creates a matrix with elements in brackets on its diagonal) and translation vectors as $T = [T_1, T_2, \cdots, T_N]^\top \in \mathbb{R}^{3N \times 1}$. Here R_n and T_n with $n = 1, 2, ..., N$ represent the rotation matrix and translation vector of the n^{th} reliant Kinect with respect to the reference.

To compute R and T, the following process can be utilised.

1) The centroid of two sets of data over all the frames and joints can be found as $A = CPD^\top / MF$ and $B = CQD^\top / MF$, where $C = [I_{3N}, I_{3N}, \cdots, I_{3N}] \in \mathbb{R}^{3N \times 3MN}$ and $D = [1, 1, \ldots, 1] \in \mathbb{R}^{1 \times F}$.
2) The error between measurement sets and their corresponding centroids for the reference Kinect and the reliant Kinects (notated as J and K respectively) can be computed as $J = P - C^\top AD$ and $K = Q - C^\top BD$, where, J and K share the same data structure with P and Q respectively.
3) A $3N \times 3N$ matrix can be computed as $H = \sum_{m=1}^{M} J_m K_m^\top$.
4) 3×3 block matrices on the diagonal of H are extracted as $L = diag\{H_{11}, H_{22}, \cdots, H_{NN}\}$, whose singular value decomposition is $SVD(L) = U \Sigma V^\top$.
5) The rotation matrix is $R = USV^\top$, where $S = I$ when $det(UV) = 1$ or $S = diag\{1, 1, -1, 1, 1, -1, \ldots, 1, 1, -1\}$ when $rank(\Sigma) = 2$ and $det(UV) = -1$ ($det\{\cdot\}$ and $rank\{\cdot\}$ find the determinant and the rank of the matrix in the brackets respectively).
6) The translation vector T can be calculated as $T = A - RB$.

III. Skeleton Data Fusion

Since our system is a linear system, it can be written as

$$x_{f+1} = Ax_f + Bw_f \quad (2)$$

$$y_f = Cx_f - \hat{T} + v_k. \quad (3)$$

Here, x_f sharing the same structure as the f^{th} column of P, except that $(X_n^m)^f = [s_1\ s_2\ s_3\ s_4\ s_5\ s_6\ s_7\ s_8\ s_9]^\top$ (the three tuples correspond to positions, velocities and accelerations along X, Y and Z directions respectively) is the estimated state of all the joints at frame f. As for y_f, it is the f^{th} column of Q. In addition, $\hat{T} = [T, T, ..., T]^\top \in \mathbb{R}^{3MN \times 1}$

Moreover

$$A = \begin{bmatrix} a & 0_9 & \cdots & 0_9 \\ 0_9 & a & \cdots & 0_9 \\ \vdots & & \ddots & \vdots \\ 0_9 & \cdots & & a \end{bmatrix} \in \mathbb{R}^{9MN \times 9MN}, \quad (4)$$

where

$$a = \begin{bmatrix} I_3 \delta t & I_3 \delta t/2 & I_3 \delta t^2/2 \\ 0_3 & I_3 \delta t & I_3 \delta t/2 \\ 0_3 & 0_3 & I_3 \delta t \end{bmatrix} \in \mathbb{R}^{9 \times 9}. \quad (5)$$

$$B = \begin{bmatrix} b & 0_{9\times 3} & \cdots & 0_{9\times 3} \\ \vdots & & & \vdots \\ 0_{9\times 3} & \cdots & & b \end{bmatrix} \in \mathbb{R}^{9MN \times 3MN}, \quad (6)$$

and

$$b = \begin{bmatrix} \delta t^2/2 & 0 & 0 \\ 0 & \delta t^2/2 & 0 \\ 0 & 0 & \delta t^2/2 \\ \delta t/2 & 0 & 0 \\ 0 & \delta t/2 & 0 \\ 0 & 0 & \delta t/2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (7)$$

where δt is the time interval between two consecutive sets of data collected from the multi-Kinect system, and $C = diag\{C_1, C_2, ..., C_M\}$, where

$$C_m = \begin{bmatrix} C_1^m & 0_{3\times 9} & \cdots & 0_{3\times 9} \\ 0_{3\times 9} & C_2^m & \cdots & 0_{3\times 9} \\ \vdots & & & \vdots \\ 0_{3\times 9} & 0_{3\times 9} & \cdots & C_N^m \end{bmatrix} \in \mathbb{R}^{3N \times 9N} \quad (8)$$

and

$$C_n^m = [R_n^{-1}\ 0_{3\times 6}] \in \mathbb{R}^{3\times 9}. \quad (9)$$

After establishing the model, Kalman filter process can be utilised to update the state in real time.

IV. Experiment and Result

To evaluate the performance of the proposed approach for multi-Kinect depth camera calibration and skeleton fusion, we performed both simulations and real-data experiments. At the same time, we validated the results of the real data experiments with the VICON system.

A. Calibration

1) Simulation: To perform the simulation in Matlab®, 3D positions of various number (1 to 5) of simulated skeleton joints generated over different length of period (1 to 10 frames) were generated and then rotated and translated with some pre-defined rotation matrices and translation vectors, which were used to simulate trajectories captured from two Kinects for calibration. To estimate the ability of this approach to counter noise, 10 db Gaussian noise was added to the rotated and translated trajectories.

	Total Frames				
	1	2	3	4	5
1	0.57/0.82	0.29/0.68	0.21/0.85	0.94/0.97	0.96/0.94
2	0.17/0.76	0.39/0.11	0.95/0.97	0.96/**0.99**	0.97/0.99
3	0.79/0.93	0.82/0.81	0.96/0.98	0.98/**0.99**	**0.99**/0.99
4	0.59/0.79	0.91/0.92	0.97/**0.99**	**0.99**/0.99	0.99/0.99
5	0.87/**0.99**	**0.99**/0.99	0.99/0.99	0.99/0.99	0.99/0.99

TABLE I: Correlation coefficients between estimated rotation matrices and translation vectors and the pre-defined ones. The first column shows the number of joints used to perform calibration

From the table, it is obvious that the accuracy of the estimated rotation matrices and translation vectors increase with time, and the more the number of joints are used, the fewer the number of total frames are needed to reach a high accurate estimation (with 0.99 correlation coefficient). As presented in Table I, when one joint is used, at least five frames are needed to compare with only two frames for five joints.

2) Real-data Experiment: In real-data experiment, we not only estimated the pose parameters between two Kinects, but also validated the data obtained from Kinects with respect to the Vicon system. Two Kinects were nonlinearly placed with their orientations towards the subject(refer to Fig. 1 for system setup). Five sets of data, including circular, front raise, helical and two random motions, with 500 frames in each performed by a healthy person were captured by two Kinects and a Vicon system simultaneously. In this experiment, we only used 3D positions of one joint (right wrist) to demonstrate that in real scenario, one joint is sufficient to calibrate two Kinects accurately. Here, due to the page limitation, we only presented the results of circular and front raise motion. In figure 2, T_{k1v}, T_{k2v} and T_{fv} represent for the rotated and translated trajetories of the raw ones from two Kinects and the fused one, while V is for raw trajectories from the Vicon system.

From figure 2, we can see that it is possible to compute a quite accurate rotation matrix and translation vector with one

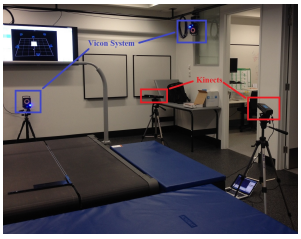
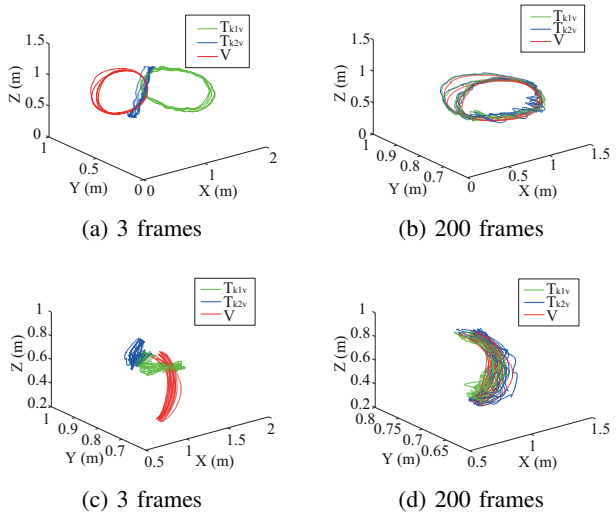


Fig. 1: System setup, including two Kinects and a Vicon system



(a) 3 frames (b) 200 frames

(c) 3 frames (d) 200 frames

Fig. 2: Rotating and translating trajectories of right wrist captured from two Kinects to the coordinate system of the Vicon system with estimated rotation matrices and translation vectors between two Kinects and the Vicon system. Fig. 2a and 2b are for circular motion and Fig. 2c and 2d are for front raise motion.

joint over approximately 200 frames from two Kinects. To validate the accuracy of this, we attached three non-linearly located markers on each Kinect so that the roll, pitch and yaw of the reliant Kinect with respect to the reference Kinect could be computed from the positions of these markers in a geometric approach, which were compared with that computed from the proposed approach. After a number of frames (around 60 for circle motion and 150 for front raise motion), the estimated relative orientation and the position of the reliant Kinect with respect to the reference one were very close ($\pm 1.75^\circ$ for the orientation and ± 0.0083, ± 0.0127 and ± 0.0109 meter for positions in X, Y and Z axis) to that computed from the positions of Vicon markers.

B. Fusion

After the calibration process, more position data (48 datasets with 500 frames in each) of right wrist from three healthy subjects was collected with this multi-Kinect system and the Vicon system. Since the sampling frequency of Kinect (30 Hz) and the Vicon system (250 Hz) were different, data captured from Kinects was re-sampled to 250 Hz. To illustrate the improvement of using this multi-Kinect system,

the raw trajectories from two Kinects and the fused one would be rotated and translated to the coordinate system of the Vicon system with previously computed rotation matrix and translation vectors between each Kinect and the Vicon system. These rotated and translated trajectories were notated as T_{k1v}, T_{k2v} and T_{fv} respectively.

Two types of statistics were carried out. Firstly, the average of root mean square errors (RMSE) between T_{k1v}, T_{k2v}, T_{fv} and the trajectory captured from the Vicon system (T_v) were computed and shown in Table II. Secondly, improvement percentages would be computed as

$$IMP_{k1f} = 100\% * (MSE_{k1v} - MSE_{fv})/MSE_{k1v},$$
$$IMP_{k2f} = 100\% * (MSE_{k2v} - MSE_{fv})/MSE_{k2v},$$

where MSE_{k1v}, MSE_{k2v} and MSE_{fv} means the MSE between T_{k1v}, T_{k2v}, T_{fv} and T_v. The means and standard deviations of the improvement percentages were computed, as well as the improvement probability, which indicates how many times the fused trajectory would be more accurate than raw ones over the total number of datasets.

	Fused (m)	**Kinect 1 (m)**	**Kinect 2 (m)**
X	0.0171	0.022	0.0241
Y	0.009	0.0116	0.0129
Z	0.0163	0.021	0.0236

TABLE II: RMSE between T_{fv}, T_{k1v}, T_{k2v} and T_v in three axes

	Mean	**STD**
IMPk1f	22.88%	18.45%
IMPk2f	27.34%	17.14%
Overall IMP	25.11%	17.80%
IP	75%	
Time	3.2860e-04	

TABLE III: Means and standard deviations (STD) of improvement percentages, as well as the improvement probability and the computational time for fusing 3D positions of one joint captured from two Kinects

From Table II, it is clear that the average RMSE of fused trajectories is smaller than those collected from the two Kinects, which illustrates the effectiveness of the fusion of skeleton data. Moreover, from Table III, it can be seen that the average improvement percentage is 25.11% (three out four times the fused trajectory is more accurate than the raw ones). Further, since the computational time for each frame is only 3.2860e-04 second, this approach can be used in real time.

V. Conclusion

Since people who need physical rehabilitation usually suffer from limited range of movement, when we deploy multi-Kinect systems to overcome the inherent issues of Kinects, we have to take their special situation into consideration. Although checkerboards have been widely used for Kinect calibration, it is not convenient for patients if they want to use a multi-Kinect system at home without caregivers. Therefore, in this paper, we proposed to use 3D positions of patient joints captured by Kinects to perform the calibration. As a result, patients are able to calibrate a multi-Kinect system easily and quickly without additional support. Moreover, since physiotherapists are likely to evaluate patients' rehabilitation exercises with their velocities, angles and angular velocities that can be easily computed from skeleton data, we proposed to skip the point-cloud-wise Kinect fusion and used skeleton data as the source and outcome of fusion directly. Although skeleton data is quite noisy and inaccurate for rehabilitation, the experiment results show that multi-Kinect system can be used to overcome this disadvantage to a certain degree. In the future, we intend to apply this multi-Kinect system in real rehabilitation environment to validate its effectiveness.

References

[1] P. D. McCann, M. E. Wootten, M. P. Kadaba, and L. U. Bigliani, "A kinematic and electromyographic study of shoulder rehabilitation exercises," *Clinical Orthopaedics and related research*, vol. 288, pp. 179–188, 1993.

[2] Kinect for windows sensor components and specifications. Access: 13 March, 2014. [Online]. Available: http://msdn.microsoft.com/en-us/library/jj131033.aspx

[3] K. Khoshelham, "Accuracy analysis of kinect depth data," in *ISPRS workshop laser scanning*, vol. 38, no. 5, 2011, p. W12.

[4] Stroke rehabilitation: Walking improves with home therapy just as well as treadmill training, study suggests. Access: 13 March, 2014, Post: February 14, 2011. [Online]. Available: http://www.sciencedaily.com/releases/2011/02/110211124617.htm

[5] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, Aug 2012, pp. 1188–1193.

[6] D. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras," *Multimedia, IEEE Transactions on*, vol. 15, no. 2, pp. 339–358, Feb 2013.

[7] E. Almazan and G. Jones, "Tracking people across multiple non-overlapping rgb-d sensors," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 831–837.

[8] R. Macknojia, A. Chavez-Aragon, P. Payeur, and R. Laganiere, "Calibration of a network of kinect sensors for robotic inspection over a large workspace," in *Robot Vision (WORV), 2013 IEEE Workshop on*, Jan 2013, pp. 184–190.

[9] M. Caon, J. Tscherrig, Y. Yue, O. Khaled, and E. Mugellini, "Extending the interaction area for view-invariant 3d gesture recognition," in *Image Processing Theory, Tools and Applications (IPTA), 2012 3rd International Conference on*, Oct 2012, pp. 293–298.

[10] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze, "Point cloud library," *IEEE Robotics & Automation Magazine*, vol. 1070, no. 9932/12, 2012.

[11] S. Asteriadis, A. Chatzitofis, D. Zarpalas, D. S. Alexiadis, and P. Daras, "Estimating human motion from multiple kinect sensors," in *Proceedings of the 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications*, ser. MIRAGE '13. New York, NY, USA: ACM, 2013, pp. 3:1–3:6. [Online]. Available: http://doi.acm.org/10.1145/2466715.2466727

[12] F. S. Kaplan, J. E. Nixon, M. Reitz, L. Rindfleish, and J. Tucker, "Age-related changes in proprioception and sensation of joint position," *Acta Orthopaedica*, vol. 56, no. 1, pp. 72–74, 1985.

[13] R. L. Gajdosik and R. W. Bohannon, "Clinical measurement of range of motion review of goniometry emphasizing reliability and validity," *Physical Therapy*, vol. 67, no. 12, pp. 1867–1872, 1987.

[14] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 4, pp. 376–380, 1991.