Automated gait analysis in people with Multiple Sclerosis using two unreferenced depth imaging sensors: Preliminary steps

S. Czarnuch[1,2] and M. Ploughman[2]
Memorial University
[1]Faculty of Engineering; [2]Faculty of Medicine
St. John's, NL

***Abstract*-Therapists often perform evaluations of gait during exercise therapies with people with multiple sclerosis. However, the level of experience, available time and characteristics of the clinical environment during the assessment makes the consistent, objective evaluation of gait challenging. Existing automated methods of gait analysis typically involve expensive equipment and can only function in specific, instrumented locations. We present the preliminary steps toward the development of an inexpensive, portable, automated gait assessment system. Specifically, we present a method of synthesizing gait data captured using two depth imaging sensors without knowledge of the location of the sensors relative to each other.**

Index Terms—Automated assessment, human gait, multiple sensors, depth sensors, point cloud.

I. Introduction

The evaluation of human gait is an important and common clinical practice [1, 2], particularly with respect to the treatment and rehabilitation of neurological diseases [3, 4], including multiple sclerosis [5]. Historically, the clinical evaluation of gait has been performed by human clinicians using both objective and subjective criteria. Objective evaluations, typically obtained using standardized tests (e.g., Timed Up and Go test), show good reliability and repeatability [6], but their sensitivity, or ability to detect clinically relevant changes in MS, has recently come under question [7]. On the other hand, subjective evaluations performed by skilled clinicians provide a rich understanding of the impact of interventions through gait observation but are not generally repeatable and require significant training and experience. Accordingly, researchers and clinicians have begun to investigate the use of more sophisticated automated techniques of gait analysis to provide more objective, sensitive and repeatable evaluations.

The gold-standard approach to automated motion analysis are three-dimensional, marker-based systems [e.g., 8, 9]. For these systems, the subject wears a specially designed tracking garment providing accurate, high-speed identification of a substantial amount of tracking points. However, these systems also require specialized, expensive equipment, require significant time to set up, and are limited to specific environments or spaces interfering with the evaluation of natural motions. Recently, marker-less approaches of human motion tracking have begun to emerge to overcome some of these limitations experienced by marker-based methodologies. Marker-less approaches, in contrast to marker-based approaches, allow actors or participants to wear their natural clothing but introduce uncertainty in measurement. These developmental approaches [e.g., 1, 10, 11], which employ RGB-D (depth) sensors, present good preliminary results, but performance of these marker-less systems has been shown to underperform when compared to the performance of marker-based systems in clinical application [12].

One of the most significant factors affecting the efficacy these marker-less approaches is the reliance on tracking data from a single depth sensor. Typical limitations of single-sensor systems are self-occlusion [12], environmental occlusion [10], single viewing angle [1], and limited sensor range [10]. Studies have employed a fronto-parallel perspective (side view) [e.g., 11], frontal perspective [13], or unique viewing angle [1, 10] to attempt to overcome the limitations of a single perspective but potentially exacerbate one limitation while addressing another.

The synthesis of data captured using multiple sensors is a promising method of overcoming these limitations. Multiple sensors may allow data that may otherwise be occluded or out of view of one (or more) sensor(s) to be captured by another sensor. Furthermore, multiple sensors can allow the system's field of view to be extended beyond that of a single sensor. To be consistent, the data from multiple sensors must first be translated into a common global coordinate space. This is accomplished through an extrinsic camera calibration procedure. This procedure requires the use of a calibration target of known proportions (generally a checkerboard cube) and the spatial location relative to the sensors is also required if the ground plane is to be calibrated. The target is used to develop transformation matrices that will allow the data from each sensor to be rotated and shifted to the global reference frame. Additionally, for the calibration to be successful, each sensor must be able to see a common surface on the calibration target to provide spatial locality. Once calibrated, the sensors can reliably transform their data to the global reference frame, allowing synthesis of the data. This procedure is not practical in clinical applications because clinicians generally do not have the skill, resources or interest to perform such technical tasks.

This paper outlines a novel approach that allows multiple arbitrarily placed sensors viewing the same scene from multiple perspectives to perform an extrinsic calibration without the use of a designated calibration target. The outcome of the method enables the data captured from the sensors to be synthesized into a single data stream in the global reference frame. This synthesis includes the knowledge of the three-dimensional rotation and offset of each sensor relative to an arbitrarily selected reference sensor, as well as knowledge of the ground plane for each sensor. The system is quasi-uninitialized, only requiring a small number of empty frames (i.e., the background scene) to be captured by each sensor.

II. Methods

To perform the extrinsic calibration, the trajectories of the centre of mass (CoM) of all objects moving in the foreground scene are used as an initial calibration target. The raw depth data captured by each sensor is first converted to a 3D point cloud using the Point Cloud Library [14]. Since the data from each sensor is captured from a different perspective, the centre of mass of each visible object in the foreground may be slightly different. Accordingly, following the initial alignment using the CoMs a final transformation to a synchronized global reference frame is performed on the raw point clouds from each sensor.

A. Experimental setup

We use a set of Microsoft Kinect RGB-D sensors connected to a single PC. The number of sensors is limited by the number of independent USB buses on the PC, where only one Kinect sensor is permitted per bus to avoid exceeding USB bandwidth. Kinect sensors were arbitrarily placed around the lab, at any elevation and perspective, ensuring that each sensor shared at least some part of the scene. An area within the lab of 5 meters by 3 meters was designated as the test scene for video capture. Video streams were captured from the Kinect sensors using the OpenNI 2.0 API [15].

Pre-trial setup: Following the placement of the sensors, a small set of 50 frames of depth data were captured simultaneously from each sensor. A background image was created for each sensor by averaging the non-zero pixels in the empty frames for each pixel $p_{avg}(i,j)$ according to the following equation:

$$p_{avg}(i,j) = \frac{\sum_{p(i,j)\neq 0} p(i,j)}{\sum_{p(i,j)\neq 0} 1} \quad (1)$$

where i and j represent the row and column of the pixel in the image similar to our previous work [16]. Only the non-zero pixels were included to remove pixels identified as *invalid* by the sensor (value of zero (0) is invalid).

Subjects: Subjects were instructed to randomly walk around the scene along to a self-directed path. The field of view of the sensors did not cover the test scene and the participants were encouraged to move outside the field of view of the sensors periodically if desired. Furthermore, the participants were not instructed to walk at any prescribed pace. Only one human was visible in the scene at any time.

B. Extrinsic calibration routine

Extrinsic calibration was performed in five main steps: 1) Foreground extraction; 2) Point cloud conversion; 3) Trajectory mapping using Centre of Mass; 4) Initial alignment; and 5) Final alignment;

Foreground extraction: The raw depth data captured by the depth sensors was first processed to remove the background. To accomplish this, depth thresholding was used, where any pixel closer to the sensor than the associated background pixel was considered part of the foreground. A threshold distance of 2cm from the background scene for each meter of distance from the sensor was used. The threshold was scaled with the distance from the sensor to account for the proportional increase in random error associated with the depth readings with increasing distance [17].

Point cloud conversion: The foreground image from each sensor was converted to a point cloud using the open-source Point Cloud Library[14]. To accomplish this, each point from the depth sensors was first converted from projective space to real-world coordinates using the intrinsic parameters (field of view and resolution) of the depth sensors. The resultant point clouds lie in the local reference frame of each sensor, without knowledge of the global reference frame or ground plane.

Trajectory mapping using of centre of mass: Procedurally, point clouds in space are generally aligned by first identifying a set of keypoints in the point cloud, then defining a set of features, and finally by aligning the keypoints and features [14]. The standard feature for a point cloud is the surface normal, largely because alignments are generally performed between two point clouds that are spatially separated by a minimal rotation and translation. The result is that the point clouds will share several common features and keypoints, allowing for effective initial alignment. We propose using the centre of mass of the objects moving in the foreground as a feature. This new feature is more suitable to our application because the sensors can be arbitrarily placed around the scene and accordingly may share little in terms of spatial structures. Without common spatial structures the initial alignment will likely fail to converge on a consistent basis.

In each new frame captured by each sensor, points were associated with a particular point cloud object if the Euclidian distance from its nearest neighbours was less than a learned threshold of 0.25cm. Point cloud objects within each frame were considered distinct if the Euclidean distance between itself and the nearest points of another object were greater than 0.25 meters. In each frame of data, the CoM was calculated for each object, and then the trajectories of the CoM of all objects were plotted over time for each sensor. A new CoM point was added to the trajectories if the object contained more the 7,500 points in its point cloud for the frame. This threshold helped ensure that spuriously detected objects (e.g., from sensor read errors) or objects almost entirely out of the scene would not skew the true CoM of the underlying object.

Initial alignment using CoM trajectories: The trajectories were used to identify an initial transformation matrix (rotation and translation) in three-dimensional space. First, a RANdom SAmple Consensus (RANSAC) method was used to fit a plane to the trajectories of each object for each sensor. This plane, in each sensor, represented a plane parallel to the ground plane the participants were walking on. Initial alignment was attempted after fifty (50) trajectory points were available from each sensor. The initial alignment was performed on the trajectories using the Sample Consensus Initial Alignment (SAC-IA) approach of [18], based on Fast Point Feature Histograms. The resulting transformation matrix was used to align (rotate and translate) the point clouds of each sensor to the ground plane of an arbitrarily-selected sensor, creating a global reference frame. The initial alignment was reattempted after ten additional trajectory points were found if the method failed to converge.

Final alignment: The initial alignment provided rough consistency between the point clouds in the global reference frame. However, the CoM's were not necessarily the actual centre of mass of the objects due to the fact that any one sensor cannot detect the full, three-dimensional profile of objects in the scene. Accordingly, some residual error, in terms of rotation and translation, was still present between the point clouds from the different sensors. A final alignment was performed using a brute-force Iterative Closest Point (ICP) method [14] on the full point clouds following the initial alignment.

III. Results

Data were captured on a laptop running 64-bit Ubuntu 14.04 LTS, with an Intel i7-350M 4-core processor. One Kinect sensor was mounted approximately parallel to the ground and approximately perpendicular to one of the walls of the lab environment. The sensor was placed at a distance of 2.5 meters from the wall to ensure the field of view was somewhat restricted (i.e., participants would periodically leave the field of view). A second Kinect sensor was mounted on top of a cabinet approximately 1.75 meters above the ground with a viewing angle approximately 60 degrees offset from the first sensor (around the normal of the ground plane), and with an approximate tilt of 20 degrees to the ground plane. The second sensor was mounted approximately 3.5 meters from the centre of the room to provide a larger field of view of the walking area and to ensure that participants periodically walked beyond the sensing range of the sensor.

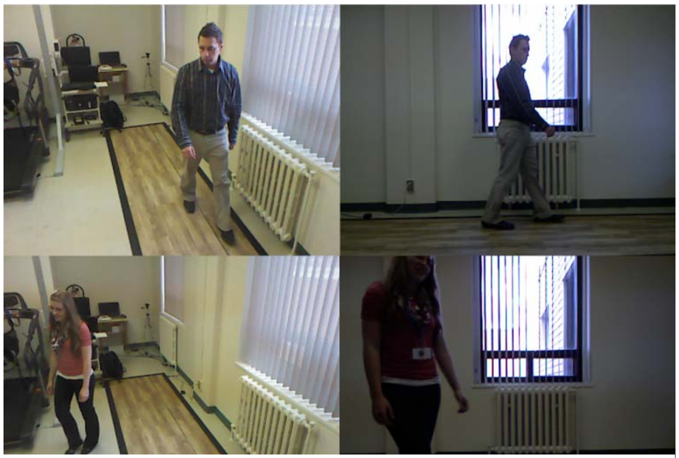


Figure 1: Sample frames captured during two walking sequences. Top row shows the participant in full view of each sensor. Bottom row shows the participant partially out of view of one sensor

Four participants contributed 1752 frames of video data, with each walking trial taking an average of 15 seconds. A total of 1455 images were included in the analyses after empty frames were removed from the beginning and end of the trials. Empty frames that were captured during the walking sequence (i.e., when the participant temporarily left the scene) were included in the analyses. A sample frame showing the scene perspective from each RGB sensor for two trial walks can be seen in Fig. 1. Images were buffered at 30 frames per second and processed after the trials were completed.

A. *Initial alignment using CoM trajectories*

A plane parallel to the ground plane for each sensor was found by implementing a RANSAC method with a minimum sample distance of 0.01 meters. The Sample Consensus Initial Alignment model was initialized with a minimum sample distance of 0.01 meters, a maximum correspondence distance of 0.10 meters, and a maximum of 1000 iterations for convergence. Across the four walking trials, the SAC-IA model converged on a solution after an average of 155 iterations.

B. *Final alignment*

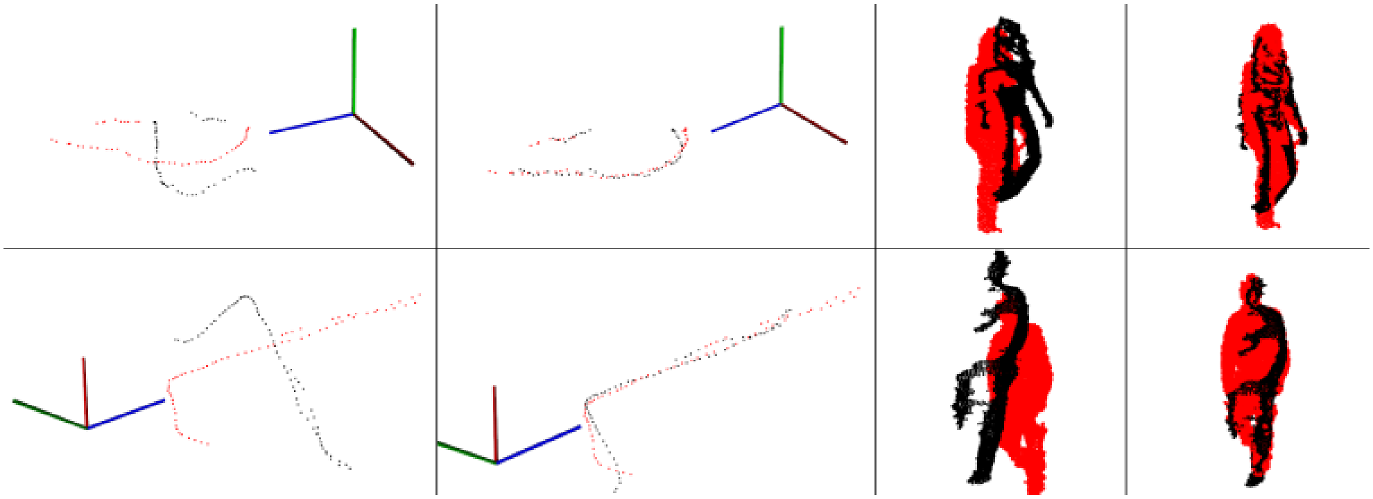
The maximum number of iterations used for the brute-force Iterative Closest Point algorithm was set to 1000. The average convergence score over the four walking trials for the ICP alignment was 0.0144, suggesting a good fit between the sensor point clouds (Table 1). Furthermore, the ICP algorithm converged after an average of 602 iterations, taking an average of 20.23 seconds. The raw and aligned trajectories, as well as the point clouds after initial and final alignment can be seen in Fig. 2.

IV. Discussion and Conclusions

The identification and coordination of a global reference frame for the synthesis of data from multiple vision sensors requires an extrinsic calibration using a calibration target. Extrinsic calibration is generally performed by a technical

TABLE I
Iterative Closest Point algorithm convergence scores and number of iterations during final alignment

	Trial				
	1	2	3	4	Mean
Convergence Score	0.00786	0.0185	0.00616	0.0249	0.0144
Iterations	574	469	631	733	602
Convergence Time	19.3s	15.8s	21.2s	24.6s	20.23s



(a) Raw trajectories before alignment

(b) Trajectories after initial alignment

(c) Point clouds after initial alignment

(d) Point clouds after final alignment

Figure 2: Trajectories and point clouds before and after alignments for two sample frames, organized by row with red representing data from sensor 0 and black representing data from sensor 1. Column (a) shows the raw centre of mass trajectories over time before alignment. Column (b) shows the trajectories after initial alignment. Column (c) shows the point clouds for each sensor after initial alignment. Column (d) shows the point clouds in the global reference frame after final alignment

expert, precluding the use of multi-sensor systems by clinicians in clinical settings. Under the assumption that the centre of mass (CoM) of objects moving in a scene can be reasonably approximated by each sensor within a multi-sensor setup, our preliminary findings suggest that the trajectories of the CoMs over time can be used to align the multiple perspectives to a global reference frame. Furthermore, our findings support that the global ground plane can also be identified using the CoM trajectories.

The preliminary success of our proposed technique of using the CoM points as both keypoints and features for the initial alignment suggests two possible advantages over the standard surface normal features in this application. Firstly, error in the identification of the CoM in each frame is negligible when the CoMs are accumulated into a temporal trajectory curve. Secondly, the need for multiple sensors to share some common spatial structures (i.e., have some overlap of the point clouds) is mitigated by using the CoMs. Performing a brute-force final alignment using an Iterative Closest Point algorithm provided good results with adequate fitness scores. Furthermore, visual inspection of the resultant point clouds in the global reference frame (see Fig. 2, column (d)) shows good coherence between the point clouds from each sensor.

A. *Limitations and future work*

This study was not without limitations. The results presented represent preliminary work toward the development of a multi-sensor system that can perform an automatic extrinsic calibration. The findings are representative of a small data sample and are partially speculative in nature. Future work will initially seek to develop more empirical methods and data to support the proposed approach.

Despite the positive results obtained from the initial alignments, the use of trajectories requires several predicate conditions to be met in order to be successful. To assure convergence, the foreground objects must move in both dimensions (relative to the ground plane). Additionally, the initial SAC-IA method seeks an alignment that minimizes the error between features (and keypoints). If the trajectories are not originally close in alignment the method may converge on the wrong solution. This situation may also occur if the trajectories are complicated (e.g., many points where the trajectory crosses itself). Note that since the method first aligns the planes of the trajectories using RANSAC, the incorrect solution will always be a two-dimensional rotational error. One promising solution to the limitations of the initial alignment method is through the use of more representative features of the CoM trajectories. The CoM points over time are more reflective of keypoints rather than features. The velocity and acceleration of these keypoints are more analogous to the standard surface normal features used on raw point clouds. Accordingly, we will continue to investigate this approach to initial alignment using the CoM points as keypoints, with the velocity and acceleration of the points as the alignment features. Furthermore, we will also consider defining more heuristic criteria for performing the initial alignment, such as limiting the size of the trajectories to simple, representative segments, ensuring motion in two dimensions is present using the RANSAC method.

The algorithms used in this study were developed to support an arbitrary number of sensors. However, this study was restricted to the use of two Kinect sensors mounted in two locations. Future work will investigate additional sensors and additional perspectives to ensure generalizability of both the initial and final alignment approaches. Additionally, the approach must be validated in different physical environments, and with multiple people walking simultaneously.

From the perspective of a system intended to perform an extrinsic calibration and run in real-time, the most significant limitation is the speed of the final alignment. Across the four trials, the average convergence time of the final alignment was 20.23 seconds. Theoretically, this alignment is only necessary a single time. However, in a real-world application, this algorithm would need to be run periodically to ensure the alignment was correct, particularly if convergence scores were not ideal. One possible solution would be to reduce the computation time by reducing the number of data points. This could be achieved by first attempting the alignment using keypoints and surface normal features. Should this significantly faster approach fail, the brute-force PCL alignment could then be performed on the full point clouds. Additionally, the alignment could be run concurrently on a separate thread (assuming adequate resources on the machine) allowing image acquisition to continue buffering images.

Notwithstanding the limitations of the current study, preliminary results are encouraging. Automatic extrinsic calibration is shown to be possible using the trajectories of the centre of mass of objects in the scene. The trajectories can be used to provide an adequate initial alignment between arbitrarily placed sensors such that a final alignment can be performed on the raw point clouds to a global reference frame. The combination of the initial and final transformation matrices theoretically allows the effective alignment of the sensors across any captured frame. Furthermore fitting a plane to the trajectories captured from each sensor using RANSAC can identify the global ground plane. These findings suggest that data from multiple arbitrarily placed sensors can be synthesized into a global reference frame with known ground plane without manual extrinsic calibration using a target calibration fixture. This automated extrinsic calibration has implications in many areas that may gain advantage from multiple sensors, and in particular may impact the automated assessment of gait in people with neurological disorders such as multiple sclerosis.

References

[1] M. Gabel, E. Renshaw, A. Schuster, and Gilad-Bachrach, "Full body gait analysis with Kinect," in *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, JP, 2012.

[2] D. Hodgins, "The importance of measuring human gait," *Medical Device Technology,* vol. 19, pp. 44-47, 2008.

[3] J. M. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton, and A. L. Goldberger, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," *Journal of Applied Physiology,* vol. 88, pp. 2045-2053, 2000.

[4] N. L. Keijsers, M. W. Horstink, and S. C. Gielen, "Ambulatory motor asessment in Parkinson's disease," *Movement Disorders,* vol. 21, pp. 34-44, 2006.

[5] M. J. Socie and J. J. Sosnoff, "Gait variability and multiple sclerosis," *Multiple Sclerosis International,* vol. 2013, p. 7, 2013.

[6] Y. Nilsagard, C. Lundholm, L.-G. Gunnarsson, and E. Denison, "Clinical relevance using timed walk tests and 'timed up and go' testing in persons with multiple sclerosis," *Physiotherapy Research International,* vol. 12, pp. 105-114, 2007.

[7] L. M. van Winsen, J. J. Kragt, E. L. Hoogervorst, C. H. Polman, and B. M. Uitdehaag, "Outcome measurement in multiple sclerosis: detection of clinically relevant improvement," *Multiple Sclerosis,* vol. 16, pp. 604-610, 2010.

[8] Vicon. (2013, September 24). *The Standard*. Available: http://www.vicon.com/standard/

[9] Qualisys. (2014, March 17). *What is motion capture?* Available: http://www.qualisys.com/

[10] E. E. Stone and M. Skubic, "Unobtrusive, continuous, in-home gait measurement using the Microsoft Kinect," *IEEE Transactions on Biomedical Engineering,* vol. 60, pp. 2925-2932, 2013.

[11] M. Ahmed, N. Al-Jawad, and A. T. Sabir, "Gait recognition based on Kinect sensor," in *SPIE 9139, Real-Time Image and Video Processing*, Brussels, Belgium, 2014.

[12] A. Pfister, A. M. West, S. Bronner, and J. A. Noah, "Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis," *Journal of Medical Engineering and Technology,* vol. 38, pp. 274-280, 2014.

[13] M. S. N. Kumar and R. V. Babu, "Human gait recognition using depth camera: A covariance based approach," presented at the Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, Mumbai, India, 2012.

[14] pointcloud.org. (2014). *Point Cloud Library*. Available: http://pointclouds.org/

[15] OpenNI. (2012, December 11). *OpenNI Modules*. Available: http://openni.org/Downloads/OpenNIModules.aspx

[16] S. Czarnuch and A. Mihailidis, "Development and evaluation of a hand tracker using depth images captured from an overhead perspective," *IEEE Journal of Biomedical and Health Informatics,* 2014 (in review).

[17] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors,* vol. 12, pp. 1437-1454, 2012.

[18] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," presented at the IEEE International Conference on Robotics and Automation, Kobe, 2009.