COMBINED MATCHING OF 2D AND 3D KINECT™ DATA TO SUPPORT INDOOR MAPPING AND NAVIGATION

Grzegorz Jóźków, Visiting Scholar
Charles Toth, Research Professor
Zoltan Koppanyi, PhD Candidate
Dorota Grejner-Brzezinska, Professor
Department of Civil, Environmental and Geodetic Engineering, The Ohio State University
470 Hitchcock Hall, 2070 Neil Avenue, Columbus, OH 43210, USA
jozkow.1@osu.edu

ABSTRACT

Small-size and low-cost imaging sensors, which are widely used in a multitude of consumer devices, provide medium-quality, typically redundant data that has the potential to be used for mapping and navigation purposes. For example, the Microsoft Kinect™ contains both passive and active imaging sensors, significantly extending the range of possibilities in its application. This study is focused on the indoor mobile use of the Kinect™ sensor for mapping and navigation. Mobile mapping of the indoor environment in 3D using Kinect™ may be realized by stitching the point cloud series acquired during movement and then reconstructing the navigation trajectory. In this study, a combined point cloud registration method is proposed that is based on 3D transformation of consecutive point clouds with transformation parameters estimated on the basis of matching 3D points. Since the modest geometrical quality of Kinect™ point clouds causes difficulties for finding corresponding 3D points, the matching was primarily performed on RGB images (2D) by applying the SIFT method. Knowing the relative orientation between Kinect™ passive and active sensors, each SIFT key-point can be associated with points in the 3D depth image and, consequently, with a 3D position. Results of several tests in a typical indoor corridor environment showed that the proposed approach for stitching point clouds allows for robust reconstruction of the trajectory and, in general, easy combination of numbers of Kinect™ image frames. Similar to other navigation systems based on image sensors alone, the Kinect™ image data also has some limitations, which are discussed in this paper.

Key words: Kinect™, image and point cloud matching, indoor navigation, 3D mapping

INTRODUCTION

Small-size and low-cost imaging sensors are widely used in a multitude of consumer devices, providing medium-quality, typically redundant data. Usual application of these devices has the potential to be extended and applied for mapping and navigation purposes. Besides typical mobile devices with imaging capabilities, such as cell phones or compact digital cameras, others such as mobile imaging sensors on vehicles and humans (personal navigation) can be used. For example, the Microsoft Kinect™ (Microsoft, 2014), which has been sold in the tens of millions, is relatively small and, although not typically mobile, can easily be carried by a human. One advantage of the Kinect™ is the availability and simultaneous use of both passive and active imaging sensors, significantly extending the possibilities of Kinect™ applications. Similar to other simple devices/sensors, the use of Kinect™ is also limited, particularly by the range of the active sensor and by data accuracy. Low data accuracy is a logical consequence of the low cost of a sensor. The influence of low data quality, however, may be somewhat offset by data redundancies. In addition, its limited range is also restrictive, being typically a few meters for the Kinect™ active sensor. This is generally acceptable for indoor mapping, especially in corridor environments where the distances between objects are generally short.

In mobile mapping, remotely sensed data are usually complemented by navigation sensor data to support platform georeferencing. In most outdoor applications, integrated GPS and IMU sensors are used for that purpose. The use of navigation sensors is not mandatory for such active sensors as laser or radar. In contrast, aerial images can be processed based just on ground control points (GCP), though the use of georeferencing is beneficial. Indoors, GPS cannot be used, posing challenges to any mapping in an unknown environment. This study aims at assessing the performance potential for indoor mapping of a low-cost sensor, Kinect™.

This work investigates using Kinect™ image sequences to support indoor navigation and mapping. Simultaneous navigation and mapping based on imagery is known as visual odometry (Scaramuzza and Fraundorfer, 2011) wherein the critical part of the computation is the matching of image frames. The Kinect™ sensor is an RGB-D camera, so the limitation of unknown scale of mono-visual odometry is not a factor. Different algorithms for frame matching in visual odometry have been proposed based on the RGB-D camera model (Huang et al., 2011; Weinmann et al., 2011; Molnar and Toth, 2013; Whelan et al., 2013; Henry et al., 2014). In this work, a relatively simple approach is proposed for Kinect™ data matching. Tests were performed with Kinect™ installed on a backpack personal navigator prototype developed earlier at the OSU Satellite Positioning and Inertial Navigation (SPIN) Laboratory (Grejner-Brzezinska et al., 2010; Toth et al., 2012; Zaydak et al., 2012). Though the use of low-cost sensors and simple algorithm may not ensure high accuracy, the Kinect™ device (or other similar RGB-D cameras) holds the potential to support indoor mapping and navigation solutions.

KINECT™ SENSOR

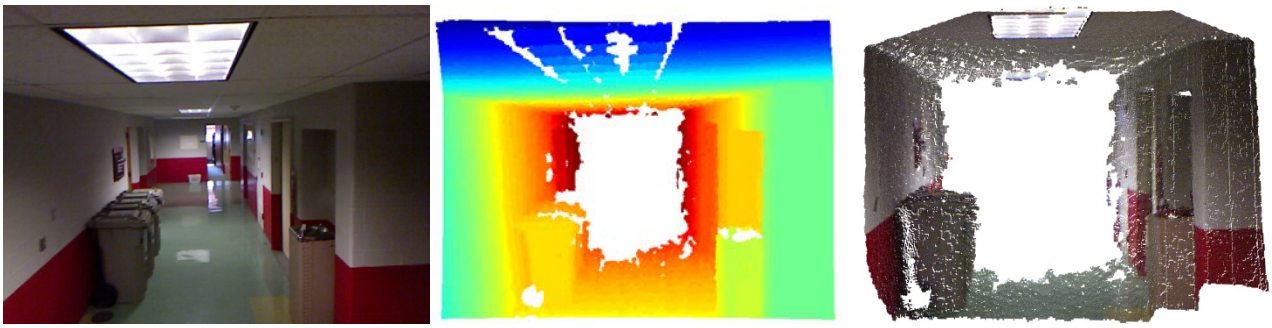
The Kinect™ sensor, the Xbox 360 video game console controller, was originally developed by PrimeSense, and later acquired by Microsoft. It allows the user to control and interact with the console just by giving voice and body gesture commands. Besides voice sensors (microphones), Kinect™ contains image sensors including a basic RGB and an IR sensor with an IR emitter (Figure 1). The emitter projects a structured light pattern of random points that is detected by the IR camera and then processed into a depth image. Detailed description of depth images generated by the Kinect™ sensor can be found in Macknojia et al. (2012). Basic parameters of the Kinect™ sensor are presented in Table 1.

Figure 1. Kinect™ sensor.

Table 1. Basic parameters of the Kinect™ sensor.

Field of view (H x V)	57° x 43°
RGB camera resolution (H x V)	640 x 480 (VGA), 1280x1024 (SXGA)
RGB camera color depth	24 bits (8 bits per channel)
IR camera resolution (H x V)	640 x 480 (interpolated) and 320 x 240
IR camera color depth	11 bits
Maximal frame rate	30 Hz

The processing algorithms as well as the firmware of Kinect™ are proprietary. For general use, Microsoft provides an SDK to support application developments. Additionally, open-source drivers are available, enabling the acquisition of raw sensor data and processing according to user's requirements. In our experiments, open-source drivers (Github, 2014) as well as open-source processing tools (OpenNI, 2014) were used. The source code was modified to preprocess data, particularly in eliminating image distortion at the data acquisition stage. Using both interior and relative orientation parameters allows for creating depth and optical images of metric quality. In our investigations, the calibration was made in earlier work (Toth et al., 2012). Based on RGB and depth images, colored point clouds can be created. Examples of RGB, depth image and colored point cloud of the same frame are shown in Figure 2. Colored point clouds are the primary data used in our investigation.



a) b) c)

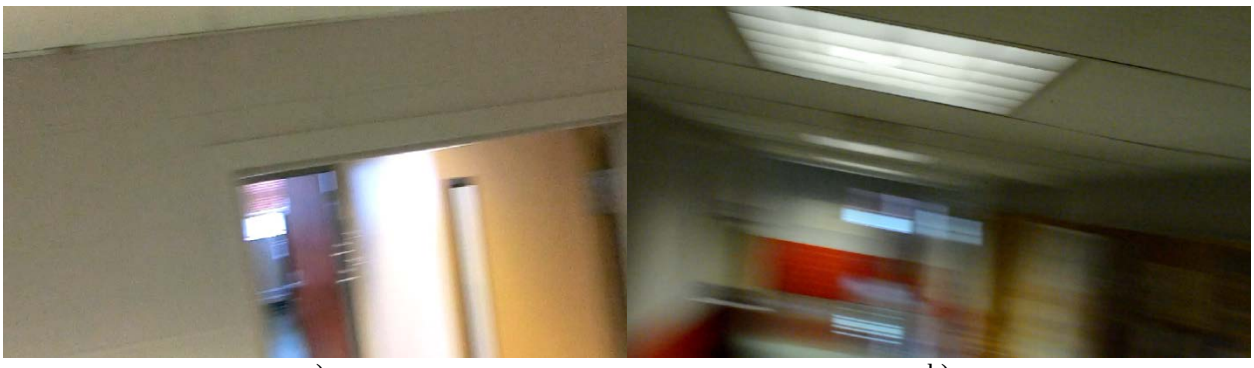
Figure 2. Kinect™ data: a) RGB image, b) depth image, and c) colored point cloud.

VISUAL ODOMENTRY USING KINECT

The approach followed in this study of indoor mapping is based on stitching the point cloud series; by reconstructing the navigation trajectory first, followed by mapping (object space reconstruction). All the processing is based on using 2D and 3D Kinect™ data. Therefore it can be seen as visual odometry. The stitched point cloud, representing the indoor mapping, was obtained in the local coordinate system which was fixed to the first frame.

Sensor Orientation

As discussed above, the primary data in our investigation are colored point clouds where color and spatial information was obtained separately by RGB and depth sensors. Considering the corridor scenario, it can be said that the front-looking sensor orientation might not be the best choice for the depth sensor as its range is limited to 10 m (Molnar and Toth, 2012), and consequently, the central part of the frame is missing. However, a side-looking orientation would result in observation of only a very small part of the mapped environment and thus would provide little diversity of ranges, which may cause problems in stitching point clouds. For the RGB sensor, the situation is reversed. The side-looking orientation provides much better geometry for spatial intersection – the principle of mono visual odometry. In the indoor environment, especially in hallways and corridors, sensor orientation perpendicular to the movement results in very close distances to mapped objects, therefore the mapped area is rather small. In addition, the short range may cause more motion blur, making the images less useful. Forward-looking camera orientation gives the best view of the mapped area, but the geometrical properties of the stereo models are the worst due to acute intersection angles. Tests with different cameras and different orientations showed that even with very low sensor platform speed, images of side/top-looking cameras are practically useless due to blur, as shown in Figure 3, and only front/back-looking orientation is suitable for obtaining appropriate RGB imagery in the indoor corridor scenario. However, it must be emphasized that conditions during the survey may change temporally; e.g., U-turns result in a situation where a forward-looking camera orientation may, for a short time, have the properties of a side-looking camera.



a) b)

Figure 3. Blur caused by side-looking camera orientation using: a) Casio Exilim EX-H20G, and b) Nikon D800.

Combined Matching of 2D and 3D Kinect™ Data

Automatic matching of point clouds can be difficult, especially for places such as corridors where the walls have almost unchanging geometry. Typically, indoor environments have more diversity in their imaging content than in their structural geometry. Thus color and/or texture information can better support the stitching together of 3D frames. The idea of combined matching that is presented in this work basically consists of first matching consecutive frames in 2D space (which is realized by image matching) and then transferring matching points onto the 3D space. Knowing the interior and relative orientation parameters of Kinect™ 2D and 3D sensors, it is possible to assign points from 2D space to matching points in 3D space and, consequently, perform point cloud registration as long as the translation and rotation parameters of each point cloud are known. The spatial relationship between two frames can be uniquely calculated based on three pairs of corresponding 3D points. Obviously, a higher number of matching points increases the reliability of the solution. However, it is not mandatory to use all possible matching points or to use techniques for dense matching that result in longer computation time, as it is more important to use points distributed as evenly as possible in the common area of the two frames. From the available image matching techniques, the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) method was selected.

The most important issue of the proposed combined 2D and 3D matching technique is the transition between 2D and 3D space. The developed Kinect™ processing software uses parameters of simultaneous calibration of RGB and IR camera, including transformation between RGB and depth images and, finally, produces colored point clouds. The 2D matching of RGB images was executed on images artificially created from the colored point clouds (Figure 4).

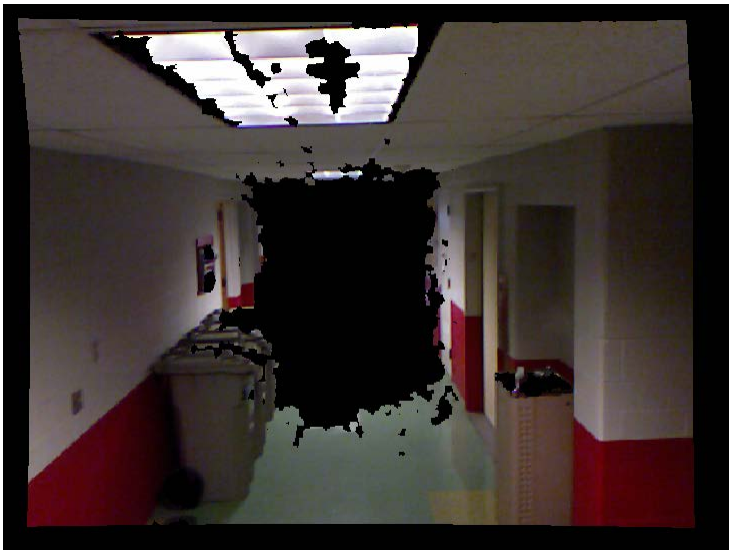


Figure 4. Example of an image created from colored point cloud.

The use of images created from point clouds was chosen for two reasons: 1) there will be no need to again calculate 3D coordinates of matched 2D points, thus the process will be faster; and 2) in places where 3D points are not present, images are blank and do not return useless feature descriptors. Creation of artificial images from point clouds is based on the pinhole projection of each 3D point to a plane that is defined by the Kinect™ sensor as: *x*-axis is the baseline of the 3D sensor, *z*-axis is the optical line of the IR camera, and *y*-axis is perpendicular to the *x*-*z* plane. Origin of the point cloud is the projection center of the IR camera. After performing SIFT and finding 2D corresponding points, there is no need to search for conjugate points in 3D space as the 3D coordinates can be treated as additional features for each pixel of the artificial image. Obviously, the final images are obtained by rounding the *x* and *y* values and translating to the image center by half of the image width. These created images are free of distortion, which was removed during data acquisition. Figure 4 illustrates the blank edges of the created images. This is caused by the different focal lengths of the two cameras: the IR camera has a somewhat larger principal distance and, therefore, a smaller field of view. Rounding error, distortion removal and the smaller field of view of the IR camera may result in multiple 3D points being projected onto the same pixel of the artificial image. For simplicity, RGB values of such pixels were averaged as well as their 3D coordinates. Additionally, the original SIFT algorithm works on a single-band image, which can be one of the RGB channels, a combined grayscale image, or an artificial intensity image (see Whelan et al., 2013). Tests showed that the distribution of SIFT keypoints may

not be sufficient to obtain acceptable results when matching is performed only on a single channel or on a grayscale image (see Figures 5a-d). Not surprisingly, the most complete distribution of matching points was obtained when results of separate SIFT matching was performed on each RGB channel, see Figure 5e. Note that adding results, obtained for the gray channel usually does not improve the distribution of corresponding points (see Figure 5f). Figures 5e and 5f summarize results after matching (note: exact repetitions of corresponding points are excluded). As uneven point distribution may negatively affect the estimation of the transformation parameters, and because there can be points very close to each other in the set of combined SIFT keypoints, a filtering mask of 5 x 5 pixels was used to remove unnecessary points.

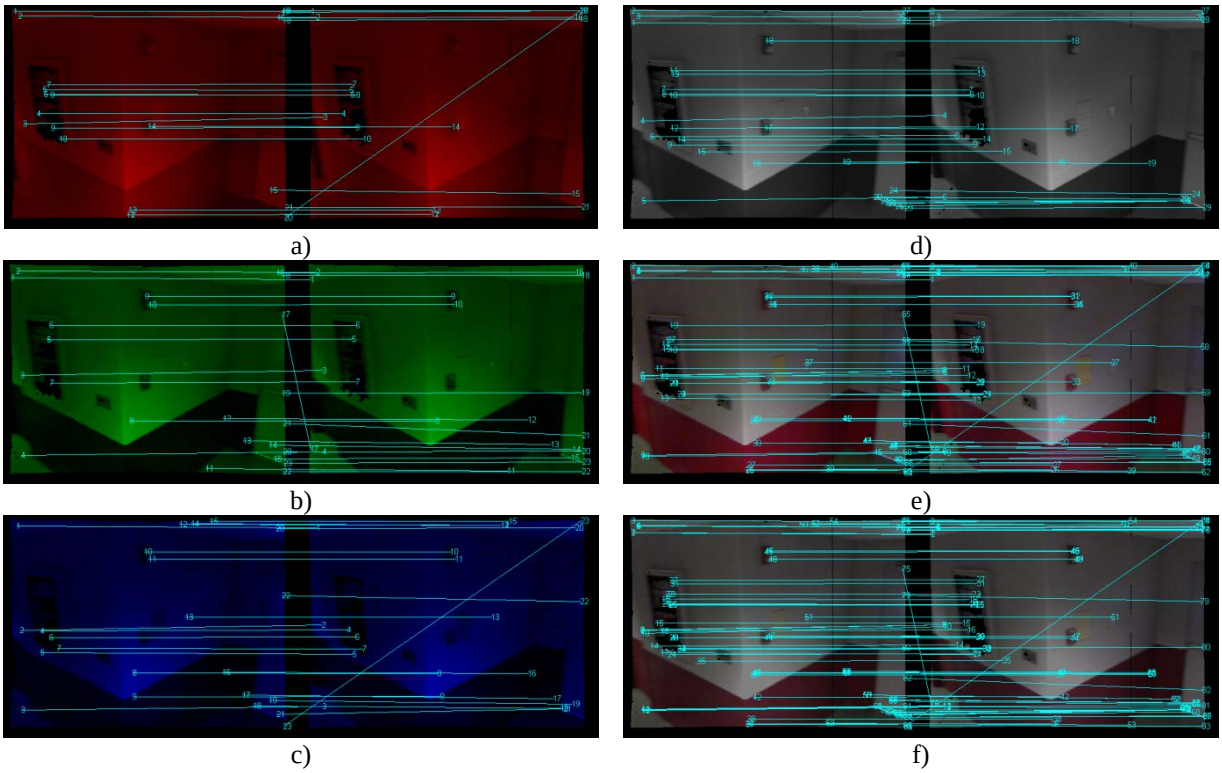


Figure 5. Distribution of SIFT keypoints for: a) R channel (21 points), b) G channel (23 points), c) B channel (23 points), d) grayscale image (29 points), e) combined RGB channels (66 points), and f) combined RGB channels and grayscale image (87 points).

Based on the 2D and subsequently 3D matching points, the transformation parameters can be estimated. When the number of matching points is much larger than the minimum number required, the least-squares method can be applied for better reliability of the estimated parameters. Since the relationship between the coordinates of 3D points and transformation parameters is non-linear, an iterative process with approximated values of estimated parameters is necessary. Approximated initial values can be calculated on the basis of a few points, though here they were assumed to be zero since consecutive frames captured at a very high rate show only small disparity. Such an assumption also makes the calculation process faster and avoids the possibility of selecting outliers as points for the initial parameter estimates. Outliers occur in 2D SIFT keypoints; they are visible in Figure 5. Incorrectly matched points have a negative impact on the estimation of transformation parameters and must be eliminated or, at least, their influence should be minimalized. In the presented work, the outliers were identified during the 3D transformation parameters estimation process. Since the number of outliers is usually smaller than the number of correctly matched points, a robust estimation can be used for eliminating wrongly matched points. In the first iteration, the robust least-squares method assumes that all points are matched correctly, and thus they all get a weight of 1. In the subsequent iterations, initial weights change based on the weighting function, and are chosen as:

$$w_i^{(k)} = \begin{cases} e^{-a\left(\left|v_i^{(k-1)}\right|-s\right)^b}, & \left|v_i^{(k-1)}\right| > s \\ 1, & otherwise \end{cases} \quad (1)$$

where:

$w_i^{(k)}$ – weight of point in *k* step of iteration,

$v_i^{(k-1)}$ – point coordinate residual in *k*-1 step of iteration, and

a, *b*, *s* – damping function parameters, empirically chosen as 1, 2, and 0.01, respectively.

Iterative processes (such as robust estimation with adjustment of initial parameters) finish if the estimated transformation parameters are not significantly different from the parameters estimated in the previous step. The value of 1e-6, radians for angles, meters for translations, was chosen as a threshold for parameter difference. Experiments showed that both iterative processes end after only a few steps, proving that the initial choices for weighting function and initial parameters were correct. Least-squares estimation of transformation parameters with elimination of outliers could be also performed using the RANdom SAmple Consensus (RANSAC) method, but the number of iterations would be much larger than that of the robust estimation method introduced here, resulting in much longer execution time. RANSAC could be used in the case where the number of outliers was larger than the number of correctly matched points, as in this case the robust estimation usually fails.

Based on the transformation parameters between consecutive frames, incremental stitching of point clouds can be performed. After transforming all frames into a common local coordinate system, a stitched point cloud is obtained.

RESULTS

Test Data

Test surveys were performed using a personal navigator backpack prototype developed earlier at the Satellite Position and Inertial Navigation Laboratory (SPIN Lab) of The Ohio State University. Mounting and facing of the Kinect™ device, as well as additional devices such as a laptop for data acquisition, are shown in Figure 6. Additional sensors mounted on the backpack prototype have been used for other investigations. Although all sensors acquired data during the tests, most of this data was not used in this work.

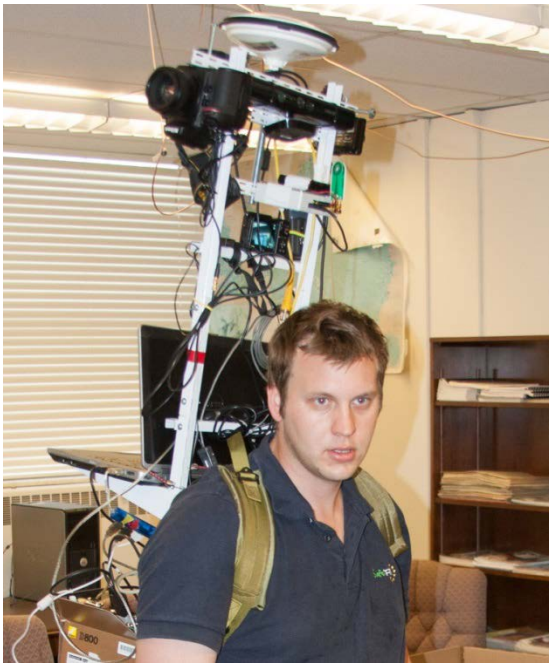


Figure 6. Personal navigation backpack prototype.

A typical office corridor scenario was chosen as the test environment. This scenario seems to be very simple for mapping and navigation, as there are straight and long halls. But, in reality it presents problems for the visual odometry approach. The lack of geometrical variety as well as color variability in plain walls and floors could make the matching of point clouds a serious challenge.

First, Kinect™ indoor surveys showed that the data acquisition frequency should be higher than 5 fps as otherwise the overlap of consecutive frames may be insufficient due to movement. Therefore, subsequent surveys with a maximum acquisition frequency of 30 fps were executed. The statistics showed that the actual frame rate still varied quite a lot (as shown in Figure 7), and that maintaining a constant acquisition speed was impossible without using a high-performance data recording system. In tests, the average acquisition speed was 26.8 fps, and over 90% of frames were captured with a rate of 20 fps or higher. Tests were executed with different human movement speeds, several U-turns and along straight and waving paths to assess the influence of these factors on the proposed trajectory reconstruction and mapping approach.

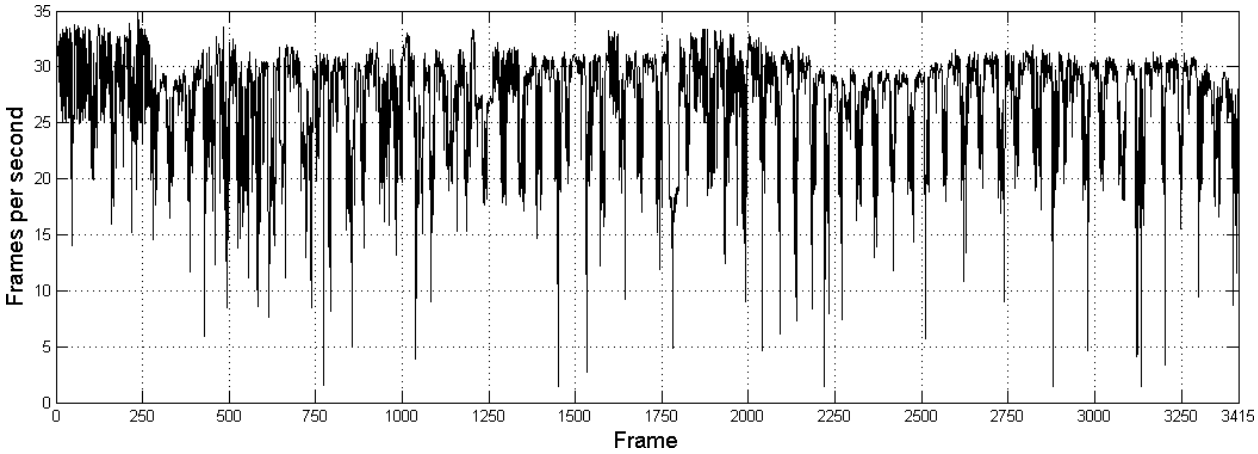


Figure 7. Empirical frame rate of Kinect™ data acquisition.

Reconstructed Trajectory and Stitched Point Cloud

As an example, the final results of about 500 stitched point clouds with reconstructed trajectory is shown in Figure 8. Clearly, the general shape of the corridor is preserved in the stitched point cloud. However, due to the incremental nature of the selected approach, some drift can be noticed. The trajectory was reconstructed quite well, though two jumps are present.

Error Identification

According to obtained results, it was noticed that most of the problems with reliable trajectory reconstruction occurred during sharp turns, usually U-turns, often related to a lower frame rate. Coincidence of these factors could result in insufficient overlap of frames and a low number of inliers in the matching point set. Additionally, during U-turns in the corridor scenario, the mapped object space is small and therefore the obtained images contain an insufficient number of features, causing inliers to be located close to each other. An example of such a situation is presented in Figure 9. Combination of a low number of unevenly distributed inliers in the set of SIFT keypoints leads to wrong transformation parameter estimation and, therefore, wrong point cloud stitching. Even the RANSAC algorithm may fail in transformation parameters estimation if correct matching points are located on a small part of the images.

To remedy these difficult situations, more complex algorithms can be considered that employ, in general, geometric features and keyframes. Although the presented algorithm is not able to avoid gross errors, they can be identified easily. Results showed that larger errors occurred for the frames with unreliably estimated transformation parameters due to speed of movement or relative to the previous frame accuracy of the trajectory points. For such types of frames, these parameters will have larger values, represented in the form of peaks in Figure 10.

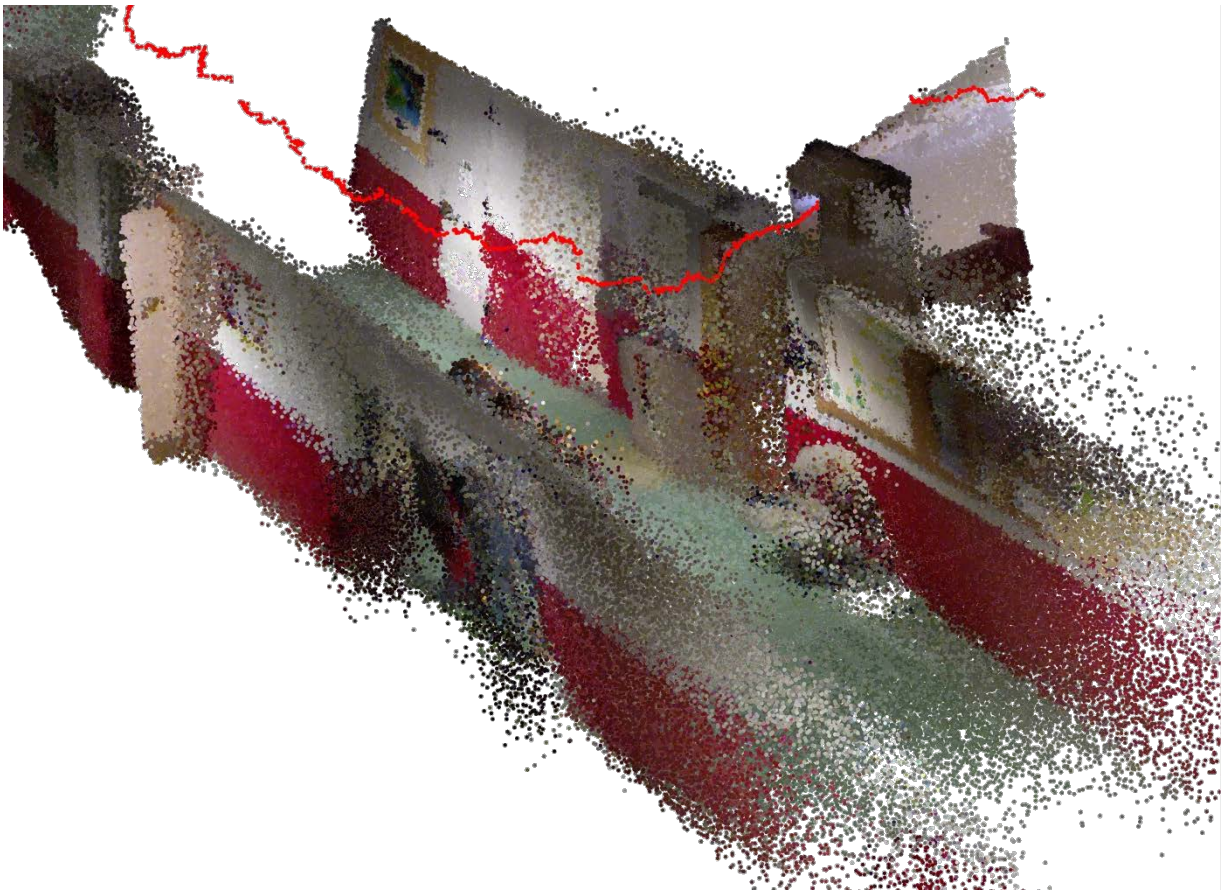


Figure 8. Stitched point cloud (ceiling removed) and reconstructed trajectory (red dots).

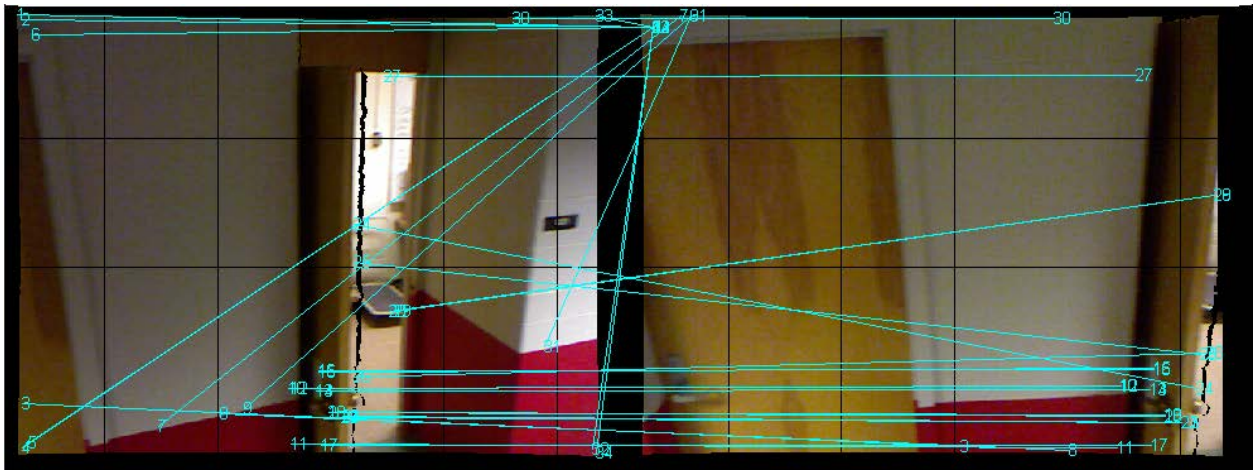


Figure 9. SIFT keypoints on the images acquired during a U-turn.

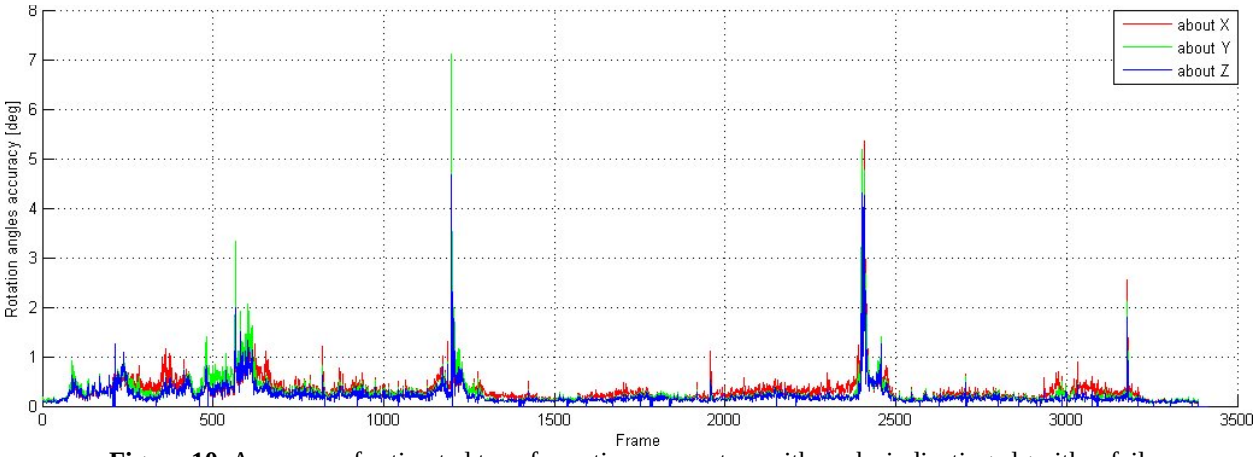


Figure 10. Accuracy of estimated transformation parameters with peaks indicating algorithm failures.

CONCLUSIONS

Initial results indicate that the method of combining Kinect™ 2D and 3D imagery for indoor navigation and mapping is feasible using low-cost RGB-D sensors. For the largest part of the survey, the obtained trajectory and stitched point cloud is a correct representation of the mapped area. The circumstances where the algorithm fails can be identified reliably. It must be emphasized that failures occurred only within specific conditions, such as U-turns, which seem to be unusual behavior during indoor navigation. Note that by introducing other sensor data, such as IMU data, these situations generally can be remedied. Finally, the proposed algorithm for point cloud stitching can be improved further by adding keyframes or implementing a Kalman filter to decrease the influence of the drift caused by a simple incremental approach.

REFERENCES

Github, 2014. PrimeSense sensor module for OpenNI, https://github.com/avin2/SensorKinect (Accessed 10 February, 2014).

Grejner-Brzezinska, D., C. Toth, J. Markiel, and S. Moafipoor, 2010. Personal navigation: Extending mobile mapping technologies into indoor environments, *Boletim De Ciencias Geodesicas, 15*(5):790-806.

Henry, P., M. Krainin, E. Herbst, X. Ren, and D. Fox, 2014. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments, In: *Experimental Robotics*, *Springer Tracts in Advanced Robotics, 79*:477-491.

Huang, A.S., A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, 2011. Visual odometry and mapping for autonomous flight using an RGB-D camera, *International Symposium on Robotics Research (ISRR)*, pp. 1-16.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision, 60*(2):91-110.

Macknojia, R., A. Chávez-Aragón, P. Payeur, and R. Laganière, 2012. Experimental characterization of two generations of Kinect's depth sensors, *IEEE International Symposium on Robotic and Sensors Environments*, pp. 150-155.

Microsoft, 2014. Kinect, http://www.xbox.com/en-us/kinect/ . (Accessed 10 February, 2014).

Molnar, B., and C. Toth, 2012. Accuracy test of Microsoft Kinect for human morphologic measurements, *XXII ISPRS Congress*, August 25, –September 01, 2012, Melbourne, Australia.

Molnar, B., and C. Toth, 2013. Spherical target based trajectory recovery from Kinect depth imagery, *Proc. ASPRS 2013 Annual Conference*, March 24-28, Baltimore, MD.

OpenNI, 2014. Open-source SDK for 3D sensors, http://www.openni.org/ . (Accessed 10 February, 2014).

Scaramuzza, D., and F. Fraundorfer, 2011. Visual odometry [tutorial], *Robotics & Automation Magazine, IEEE, 18*(4):80-92.

Toth, C., B. Molnar, B., A. Zaydak, and D. Grejner-Brzezinska, 2012. Calibrating the MS Kinect sensor, *Proc. ASPRS 2012 Annual Conference*, Sacramento, USA.

Weinmann, M., S. Wursthorn, and B. Jutzi, 2011. Semi-automatic image-based co-registration of range imaging data with different characteristics, *ISPRS Journal of Photogrammetry and Remote Sensing*, *38*(3):W22, pp. 119-124.

Whelan, T., H. Johannsson, M. Kaess, J.J. Leonard, and J. McDonald, 2013. Robust real-time visual odometry for dense RGB-D mapping, *2013 IEEE International Conference on Robotics and Automation*, pp. 5724-5731.

Zaydak, A., C. Toth, D. Grejner-Brzezinska, B. Molnar, Y. Yi, and J.N. Markiel, 2012. 3D image-based Navigation in collaborative navigation environment, *Proc. ION GNSS*, September 17-21, 2012, Nashville, pp. 2462-2468, CD ROM.