Comparative study of RGB-D sensors based on controlled displacements of a ground-truth model

Adrien Anxionnat, Sandrine Voros, Jocelyne Troccaz

Comparative study of RGB-D sensors based on controlled displacements of a ground-truth model[1]

Adrien Anxionnat [a], Sandrine Voros [a], and Jocelyne Troccaz [a], *Fellow, IEEE*

***Abstract*— In the context of developing a pedagogical tool for teaching anatomy, the need for a comparative study between two RGB-D cameras has emerged. This paper addresses the assessment of the accuracy and precision of two RGB-D sensors (Carmine Primesense and Persee Orbbec) through two different experiments. The evaluation not only provides comparative results on the sensors performances but also aims at determining in which conditions they are the most efficient. The first experiment evaluates the variability of the output depth map data. The second experiment focuses on analyzing the influence of the distance in the positioning accuracy of an object submitted to controlled displacements. The results are summarized in a set of error heat maps and a table; they provide clues for using one sensor rather than another by describing their robustness both in a static scene and in a motion capture scenario.**

I. Introduction

This work has been conducted in the context of a collaborative research project which objective is to propose innovative pedagogical tools for teaching of anatomy. This project has already resulted in the "Living Book of Anatomy", a mirror-like augmented reality (AR) system [1] enabling a trainee to see his/her anatomy in motion super-imposed to his/her image captured by a single RGB-D camera. In order to further develop the system through the addition of other cameras for capturing a larger posture range, we studied a new RGB-D camera, the Orbbec Persee (2016), and compared it to an older one, the Primesense Carmine (2011) which was initially used in the project.

Since its commercialization in 2010, the Kinect sensor has gained a lot of momentum in biomedicine. Not only is it a portable and non-invasive device, it is also an affordable way of tracking movements, recognizing objects and modeling scenes. Recently, RGB-D cameras were used to perform fall detection of elderly people with multi-modal features including color images and skeleton data [2] or to assess the dementia disease degree with recurrent neural network [3]. It was also involved in the design of augmented reality systems for applications in anatomy education [1], [4].

Some methods have been investigated in order to assess and to compare the performance of RGB-D sensors. In 2012, a complete study was conducted in order to deeply investigate the accuracy and precision of the Microsoft Kinect device [5]. The accuracy of the sensor was determined by comparing a scene depth map of a Kinect sensor with that of a laser camera. The laser camera was considered as ground-truth data. The two resulting point clouds were registered in order to be comparable. Two different registration methods were used: one based on manual initialization and Iterative Closest Point (ICP) and the other one on RANSAC algorithm. The precision of the sensor was assessed through the study of the standard-deviation of the depth measure error as a function of the distance to the sensor. To achieve this, a plane was fitted on a door surface at different distances and the residual error on the region of interest was calculated. In [6], Haggag et al. studied the resolution of the sensor by computing, for various plane-camera distances, the smallest discrepancy in the retrieved depth maps. They also evaluated the entropy of depth measurement for each pixel of a static scene.

One requirement of our collaborative research project was to capture the motion of a human being from one frame to the other. Although [5] carried out an interesting experiment on systematic errors of RGB-D cameras, their method depends on the registration between the camera data and other modality data, which may be a source of error. The study of entropy in [6] does not give any quantitative insight about the camera precision in terms of distance, which is a crucial parameter in our project. Moreover, neither [5] nor [6] studied the random errors for a more complex object than a plane, although a complex geometry exhibits common issues that might occur in routine usage, in particular lighting changes and partially hidden zones.

The purpose of the present study is to propose new methods for assessing the performance of RGB-D sensors without using any other modality. We designed two experiments for evaluating the capability of a camera to retrieve a known transformation of an object between two frames. The first one gives quantitative information on the measurement precision of the two cameras. The second experiment compares known motions of this reference object with the motions computed from the RGB-D sensor measurements.

II. Material

RGB-D devices are composed of an RGB camera, an IR projector and an IR sensor. An IR pattern is projected by the camera, the deformation of this pattern is captured by the sensor and analyzed in order to create a depth map of the scene. The two cameras studied in this work are the Primesense

This work was partly supported by the French ANR within the "Investissements d'Avenir" program (Labex CAMI ANR-11-LABX) and through the An@tomy2020project (ANR-16-CE38-0011).

Authors are from :

[a] Univ. Grenoble Alpes, CNRS, INSERM, Grenoble INP, TIMC-IMAG, F-38000 Grenoble, France (email: firstname.lastname@univ-grenoble-alpes.fr).

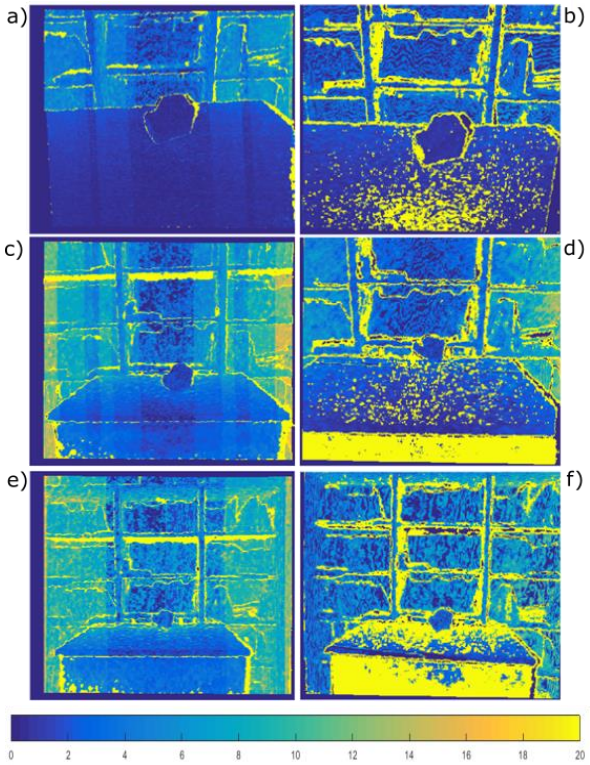

Figure 1. Heat map of the standard-deviation of depth values for each pixel (in millimeters) computed over 100 images for three different distances: 800 mm, 1300 mm and 1800 mm (**up to down**) and for two different sensors: Primesense Carmine sensor data (**a, c, e**) and Orbbec Persee data (**b,d,f**). Deep blue corresponds to a 0-mm-standard-deviation and every values greater than or equal to 20 mm of standard deviation appear in yellow.

Carmine (2011) and the Orbbec Persee (2016). Both of them provide 480x640 depth map frames at 30 FPS (Frames per Second). The range of Primesense Carmine is 0.8 m to 3.5 m and the one of Orbbec Persee is 0.6 m to 8 m.

For the experiments, we have designed and printed a 3D reference object (a polyhedron consisting of eight faces, see Fig. 2). The geometry of the object is not representative of the human body, but its structure is complex enough to assess and compare the RGB-D cameras. The object size is 15 cm x 10 cm x 10 cm, it is adapted to the camera field of view at the studied ranges. It has been printed by a Makerbot Replicator 2X 3D printer that features a 100-micron layer-height resolution. It was made of PLA (polyactic acid) and was painted in white for better reflectance of IR. For the second experiment, the polyhedron is represented by a cloud of 4000 points in the object reference frame.

III. Precision Evaluation

The idea of the first experiment was to capture the reference object and to study the variability of each pixel of the depth map. It aimed at describing the repeatability of the two sensors especially regarding the quality of the point cloud representing the reference object.

A. Method

Let us note the depth image as followed:

$$I_D(u,v) = d$$

With:

- (u,v), the pixel coordinates.
- d, the distance from the camera origin to the scene at pixel (u,v).

Each RGB-D camera is placed successively at three different distances from the object: 800 mm, 1300 mm and 1800 mm. For each distance, 100 depth maps of a static scene featuring the reference object are acquired. Standard deviation is computed independently for each pixel among the 100 samples.

$$I_{SD}(u,v) = \sqrt{\frac{1}{100}\sum_{i=1}^{100}\left(I_D^{(i)}(u,v) - I_{mean}(u,v)\right)^2}$$

B. Results

Results are shown in Fig. 1. These tests permitted to make sure the reference object was reasonably well represented in the depth maps. It provides an initial view of variability of depth measure on each device, the standard deviation being an indicator of the data consistency. The depth variability on edges for both sensors is significantly high. In particular, the edges separating the background from the reference object are not precisely captured. More generally, on the areas that are not occluded, the standard deviation is between 0 mm and 10 mm, this is consistent with the study in [6]. One can notice that Persee data feature generally more noise than Carmine data. The metallic texture of the table causes errors for Orbbec Persee, probably due to disturbed reflectance of the IR pattern. Some vertical bands can be seen on Primesense data as well.

IV. Accuracy

The objective of the second experiment is to assess and compare the accuracy of the two RGB-D sensors in a controlled experiment with a ground truth measurement table. The reference object is translated with the table and its displacements are measured with the two RGB-D sensors. The measured transformation is compared to the ground-truth transformation given by the table. A measurement table (Zaber, Model LSQ-600B-E01-T3) is used in order to provide the ground truth in the motion analysis. It can move a fixed object with high precision (0.49 µm per motor step) in two directions in a range of 60x60 cm². Fig. 2 illustrates the basic idea of the setup.

A. Method

The reference object is fixed on the table and is illuminated by the IR projector of an RGB-D camera. The object is moved of 100 mm on a grid of size 6 x 6. The distance to the sensor ranges from 800 to 1300 mm. On each transverse line, the acquisitions are performed four times in order to have a better

statistic on the data. This gives a total of N=144 (24x6) measurements of translations of 100 mm. Because the position of the camera relative to the table is unknown, we study the norm of the object motion instead of the motion itself. For each of the translations, the error between the norm of the computed translation and the norm of the ground truth displacement of the table is computed. The error mean and standard deviation are computed for each line in order to estimate the impact of the object distance on the accuracy of the sensor.

$$Err_{mean} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|{}^{C}t_{mes}^{(i)}\right\| - \left\|{}^{T}t_{ref}^{(i)}\right\|\right)$$

$$Err_{var} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|{}^{C}t_{mes}^{(i)}\right\| - \left\|{}^{T}t_{ref}^{(i)}\right\|\right)^{2} - (Err_{mean})^{2}$$

With:

- ${}^{C}t_{mes}^{(i)}$, the translation measured in the camera frame C.
- ${}^{T}t_{ref}^{(i)}$, the ground truth translation expressed in the table frame T.

The vector ${}^{C}t_{mes}^{(i)}$ is found by computing the transformation matrix ${}^{C}_{M}\hat{T}$ from the model frame (noted M) to the camera frame (noted C) for two successive positions of the calibrating model (Fig. 2). In this formalism, one can express ${}^{C}t_{mes}^{(i)}$ as follows:

$${}^{C}t_{mes}^{(i)} = {}^{C}_{M}\hat{t}^{(i+1)} - {}^{C}_{M}\hat{t}^{(i)} \qquad 1.$$

${}^{C}_{M}\hat{T}^{(i)}$ is computed using the ICP method [7]: it fits the object model point cloud to the object in the captured scene at each acquisition. It is based on a coarse-to-fine implementation of the ICP algorithm which improves the matching performance [8]. It takes as input the scene and model point clouds and it triggers as output the refined transformation between both set of points. The algorithm iterates over different scales (from coarse-to-fine) and for each scale over a predefined number of iterations.

Each iteration of the algorithm can be divided in four steps:

1) *Compute closest points.*
2) *Weigh the couplings.*
3) *Find the optimal rigid transformation.*
4) *Apply the transformation to the model.*

The algorithm stops either if the number of iterations has reached a maximum or if the optimal rigid transformation of the current iteration is close enough to identity according to a defined parameter.

The scale parameter of the ICP is fixed to 6. Tolerance and rejection scale parameters of the ICP have been set (respectively 0.06 and 0.4) in order to reject easily aberrant matching. The maximum number of iterations has been set to 300.

Because the ICP needs a robust initialization, a paired-point matching method ("Arun" [9]) is employed with manually-selected points. More precisely, a region of interest is set by the user in order to crop the scene point cloud around the object which has to be registered to the model. Then, the N_A most

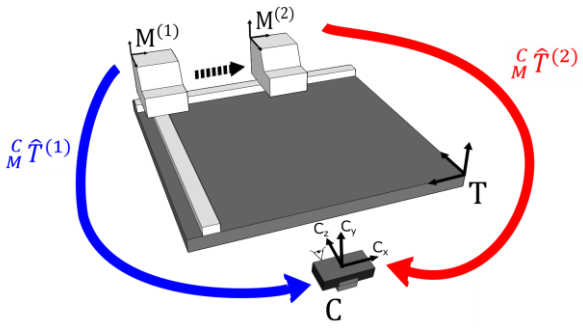

Figure 2. Illustration of the relationship between two successive acquisitions involved in the experiment on a transverse line. Frames and transformations involved.

visible vertices p_i^s of the object points cloud are matched by the user to corresponding points p_i^m in the reference model frame (most often N_A =3). The optimal rotation R_A and translation t_A are computed such that they minimize the functional in Eq. 2:

$$\arg\min_{R_A\,T_A}\sum_{i=1}^{N_A}\|p_i^s - R_A * p_i^m - t_A\|^2 \qquad 2.$$

With:

- p_i^S and p_i^M, the N_A corresponding column points between the scene and the model.
- R_A, the optimal rotation computed after manual matching.
- t_A, the optimal translation computed after manual matching.

The homogeneous transform computed with the ICP transformation, T_{ICP}, is applied to the initialized model point cloud $T_A * P^m$:

$$P_{final}^m = T_{ICP} * T_A * P^m$$

Where:

- $P^m = \begin{bmatrix} x_1^m & \dots & x_{N^m}^m \\ y_1^m & \dots & y_{N^m}^m \\ z_1^m & \dots & z_{N^m}^m \\ 1 & \dots & 1 \end{bmatrix}$, the matrix of homogeneous coordinates of the model point cloud before transformation.
- $T_A = \begin{bmatrix} & R_A & & t_A \\ 0 & 0 & 0 & 1 \end{bmatrix}$, the initial homogeneous transform computed with Arun method.
- T_{ICP} is the homogeneous transform computed with multi-scale ICP method.

At the end of the procedure, the position of each point of the model is known in the camera frame. The estimate of the transformation matrix from the frame of the model M to the frame of the camera C, written ${}^{C}_{M}\hat{T}$ can be expressed as:

$$ {}^{C}_{M}\hat{T} = T_{ICP} * T_A = \begin{bmatrix} R_{ICP} * R_A & R_{ICP} * t_A + t_{ICP} \\ 0 \quad 0 \quad 0 & 1 \end{bmatrix} $$

It is hence possible to compute ${}^{C}t^{(i)}_{mes}$ defined in Eq. 1:

${}^{C}t^{(i)}_{mes} = (R_{ICP} * t_A + t_{ICP})^{(i+1)} - (R_{ICP} * t_A + t_{ICP})^{(i)}$ and finally to get the mean error Err_{mean} and the variance Err_{var} (over all the translations) as a function of the six distances from the sensor.

B. Results

The results are presented in Table. 1. The Carmine sensor systematically under-evaluates the motion by 6 to 8 millimeters for the considered distances. The Persee device is accurate at the lowest distance (800 mm) but its performance decreases of around 4% at 900 mm and stagnates over 900mm. The standard-deviation for Primesense is generally low compared to Orbbec.

V. Discussion

The proposed study compares the precision and accuracy of two RGB-D sensors (Primesense Carmine and Orbbec Persee). The first experiment provides comparative clues on the variability of both sensors in a static scene for different distances. The second experiment aimed at measuring the accuracy of each sensor by comparing translations of an object measured with each sensor to ground-truth translations. We can conclude that Primesense Carmine has better performance than Orbbec Persee in terms of precision. The systematic shift suggests that an additional calibration could improve this performance. The high values for Orbbec standard deviation at 1300 mm seem mainly due to difficulties to perform accurate registration in presence of very noisy data (cf. first experiment) and when the object is placed far from the sensor. It might be problematic when motion must be estimated with high accuracy. In the context of our project, we plan to use several cameras in order to capture a given scene from different points of view. We will therefore position our cameras regarding their performance i.e. we will prioritize the use of Persee at close range.

The choice of a suitable camera is a crucial step in the design of experimental protocols, and this paper describes a methodology to evaluate the performance of two affordable cameras on the market thanks to the use of a 3D printer and a measurement table, which could be applied to any RGB-D camera. Future works include the fusion of data of two or three RGB-D sensors in order to improve the performance of pose estimation. Another research direction is to assess the capability of dynamic capture for different sensors.

TABLE I. Comparative study of two RGB-D Cameras

Distance (mm)	**Mean (standard-deviation) of errors in millimeters**	
	Primesense Carmine	***Orbbec Persee***
800	-6.94 (0.90)	1.72 (3.32)
900	-6.96 (0.91)	5.82 (3.86)
1000	-7.28 (0.71)	4.99 (4.99)
1100	-7.57 (0.74)	4.25 (3.76)
1200	-7.94 (0.69)	5.57 (4.50)
1300	- 8.12 (0.99)	11.25 (16.62)

Acknowledgment

We thank the ECCAMI platform (http://www.eccami.fr/) for giving us access to the XY precision table.

References

[1] A. Bauer *et al.*, "Anatomical augmented reality with 3D commodity tracking and image-space alignment," *Comput. Graph.*, vol. 69, pp. 140–153, Dec. 2017.

[2] T.-H. Tran, T.-L. Le, V.-N. Hoang, and H. Vu, "Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment," *Comput. Methods Programs Biomed.*, vol. 146, pp. 151–165, Jul. 2017.

[3] S. Iarlori, F. Ferracuti, A. Giantomassi, and S. Longhi, "RGB-D Video Monitoring System to Assess the Dementia Disease State Based on Recurrent Neural Networks with Parametric Bias Action Recognition and DAFS Index Evaluation," in *Computers Helping People with Special Needs*, 2014, pp. 156–163.

[4] M. Meng *et al.*, "Kinect for interactive AR anatomy learning," in *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013, pp. 277–278.

[5] K. Khoshelham and S. O. Elberink, "Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, Feb. 2012.

[6] H. Haggag, M. Hossny, D. Filippidis, D. Creighton, S. Nahavandi, and V. Puri, "Measuring depth accuracy in RGBD cameras," in *2013, 7th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2013, pp. 1–7.

[7] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[8] T. Jost and H. Hugli, "A multi-resolution ICP with heuristic closest point search for fast and robust 3D registration of range images," in *Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings.*, 2003, pp. 427–433.

[9] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 698–700, Sep. 1987.