Volume measurement with a consumer depth camera based on structured infrared light

Babette Dellen [1] and Ivan Rojas
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Llorens i Artigas 4-6, 08028 Barcelona, Spain

Abstract. The measurement of object volumes is of large importance for many sectors in industry, including agriculture, transportation, production, and forestry. In this paper, we investigate the feasibility of using commercial depth-sensing devices based on structured light such as the Kinect camera for volume measurement of objects of medium size. Using a fixed set-up, depth data are acquired for different views of the object and merged. Volumes are carved using a volume-intersection approach, which is computationally simple, and, most importantly, model-free. The performance of the method is evaluated using ground-truth volumes of a benchmark data set of selected objects, and volume-measurement errors are reported for a set of household objects.

Keywords. volume measurement, 3D reconstruction, volume intersection, depth camera

1. Introduction

Non-destructive measurement of object volumes is required in many areas of industry [1, 2]. For example, in agriculture, horticultural products need to be graded based on size and weight. In transportation, the sizes of parcels and pallets need to be estimated in order to calculate shipping costs. To meet this need, various electronic systems have been developed over the past few decades, among them are three-dimensional machine vision systems based on active methods [3, 4, 5]. Solutions have been mostly developed for specific applications and products [6, 7, 8]. For example, in [8] a system for the measurement of oyster meat volumes based on laser triangulation was proposed, where the volume was estimated from height variations in laser scan lines. In another work, the 3D shape of tomato fruits was reconstructed using a laser range finder for fruit quality classification [6]. There are also commercial systems for volume measurement available on the market, mostly for parcels and pallets. The latter systems however often do not pay off for companies with a small warehouse having to process only a few shipments a day. This motivated us to investigate the possibility of using low-cost consumer depth cameras for estimating object volumes.

[1]Corresponding Author: Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Llorens i Artigas 4-6, Barcelona, Spain; E-mail:bdellen@iri.upc.edu

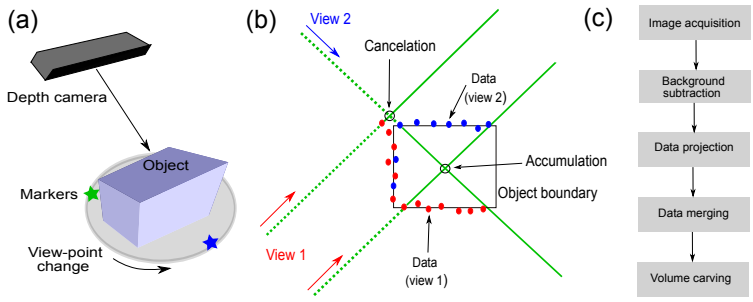


Figure 1. Overview of the procedure: (a) Set-up consisting of a turn table, on which the object is placed, a Kinect camera, and markers. (b) Schematic illustrating the volume-intersection method. Data points from two different views of the object (circles) define parts of the silhouette of the object (black lines). Back projection of the data points for each view leads to an accumulation of their respective contributions inside the object boundary, while forward projections, counting negatively, lead to their cancelation outside the object boundary. (c) Flow diagram showing the main steps of the procedure.

Recently, the release of the Kinect camera (`www.xbox.com/en-US/kinect`) [2], a depth sensor based on a structured infrared-light system, has opened new possibilities for acquiring depth information. It has a ranging limit of roughly 0.7 to 6 m distance, and is applicable in most indoor environments. Experimental results have shown that the random error of depth measurement increases with increasing distance to the sensor, and ranges from a few millimeters up to about 4 cm at the maximum range of the sensor [9]. This makes the Kinect potentially a useful device for measuring the volume of medium sized objects such as parcels and alike. Since its release, the Kinect camera has been used in many works, among them methods for the measurement of 3D structures and scene reconstruction [10, 11].

A popular way for obtaining the shape of objects are volume-intersection methods [12, 13, 14]. Object silhouettes from different views of the object can be used to find a bounding volume that is formed by back-projecting the silhouettes. Volume intersection methods have the advantage that they can be applied in situations where the acquired data (i) does not only contain surface data, (ii) is non-uniformly distributed, (iii) contains noise, (iv) and no *a priori* knowledge of the object's shape is available [13]. For many volume measurement tasks, such as the estimation of shipping costs of parcels, the resolution of small object details and concavities is not important, making volume intersection an adequate choice for this task.

In this work, we apply a volume-intersection approach to the scattered 3D object data acquired with the Kinect using a fixed set-up with known coordinate transformations between views. We further apply back-projection of silhouette points before transforming the data into a common coordinate system, which simplifies the computational aspect of this step, and include forward-projection of silhouette points, counting negatively during volume carving, which helps reducing errors caused by faulty data points, due to either measurement noise or imperfect merging of point clouds near the true object boundary. The estimated volumes are compared with hand-measured volumes for a

[2] Trade and company names are included for benefit of the reader and imply no endorsement or preferential treatment of the product by the authors

Figure 2. Pictures of typical objects used in the experiments.

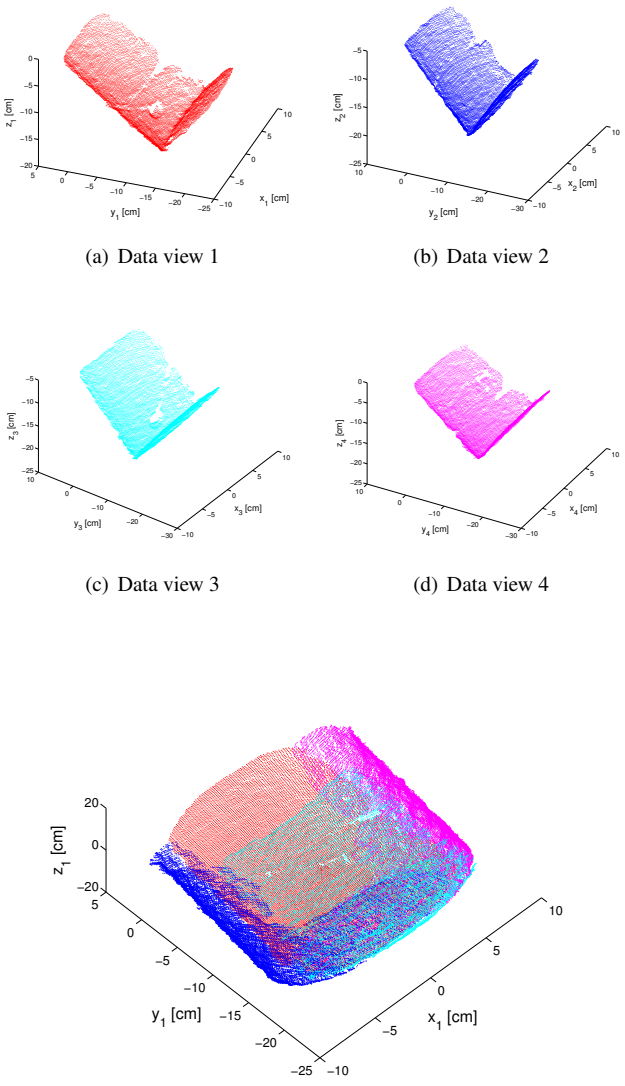
set of domestic objects having different shapes and sizes, and the feasibility of using the Kinect for basic volume-measurement tasks is evaluated.

2. Method

The procedure for volume measurement consists of the following, consecutive steps. First, images for four different views of the object are acquired with the Kinect (see Section 2.1 and Fig. 1(a)). From the images, point clouds are extracted, and the background is subtracted. For each view, we extend the data using back and forward projections, providing several sets of data points (Section 2.3). After this step, the data sets of the different views are merged by transforming them to a common coordinate system (see Section 2.2). By discretization of the 3D space, an accumulation matrix is defined, which is used to sum the contributions from the back and forward projections of the data points. Contributions from back-projected points are assigned a positive value, while the ones from forward-projected points are counted negatively. Values then accumulate positively in the area of the object in the accumulation matrix (see Fig. 1(b)). To fill holes in the data, the accumulation matrix is smoothed. The 3D area of the object is found by thresholding the accumulation matrix. Then, the volume is calculated as the sum of its discrete elements (see Section 2.4). A flow diagram of the procedure is provided in Fig. 1(c).

2.1. Set-up and image acquisition

First, objects are placed on a turn table. The Kinect camera is placed at distance of about 80 cm from the object. Depth images are acquired for four different positions of the turn table, resulting in different views of the object. Note that rotating the turn table can be understood as a view-point change from the camera's point of view. The angular position of the turn table is changed in steps of 90 degrees by rotating the table around its center. A schematic is provided in Fig. 1(a). Using the PCL library



(a) Data view 1

(b) Data view 2

(c) Data view 3

(d) Data view 4

(e) Merged views

Figure 3. View merging: (a)-(d) Points clouds after background subtraction for four different views. (e) Merged point clouds presented in the coordinate system of view 1.

(`http://pointclouds.org`), the point clouds for each view are extracted. Data points belonging to the background (floor) are removed by fitting a plane using the RANSAC algorithm (PCL library) and applying a distance threshold.

2.2. Estimation of transformation parameters

By attaching four markers to the turn table (one *out* of the base plane of the turn table), point correspondences between different views, generated by rotating the turn table in steps of 90 degrees, are established. Using these correspondences, the linear transformation M_j of the current view v_j with the first view is found using Procrustes analysis. Once the transformation matrices for the different positions of the turn table are known, they can be used to merge the object data that have been acquired using the same positions of the turn table.

2.3. Data projection

Before merging, the data for a given view is provided in the camera coordinate system, hence, the viewing direction points in the z-direction of the coordinate system. For each view v_j and data point (x_i, y_i, z_i), a set of back-projected points

$$P_{i,j}^{\text{back}} = \{(x,y,z) : x = x_i,\, y = y_i, \text{ and } z > z_i\} \tag{1}$$

and a set of forward-projected points

$$P_{i,j}^{\text{forward}} = \{(x,y,z) : x = x_i,\, y = y_i, \text{ and } z < z_i\} \quad , \tag{2}$$

is created, as illustrated Fig. 1(b). Points are generated at regular distances, which are chosen to be equal to the length d of the unit cubes used for defining the accumulation matrix in Section 2.4. Finally, we define $P_j^{\text{back}} = \bigcup_i P_{i,j}^{\text{back}}$ and $P_j^{\text{forward}} = \bigcup_i P_{i,j}^{\text{forward}}$ for each view.

When transforming the data points into a common coordinate system (here the coordinate system of the first view), we distinguish between the back- and forward-projected points, because they have to enter the computations for volume carving differently.

2.4. Data merging and volume carving

The merging of the data is achieved by applying the previously found transformation between the views to every 3D point in P_j^{back} and P_j^{forward}, yielding for each view two sets of transformed points $\tilde{P}_j^{\text{back}}$ and $\tilde{P}_j^{\text{forward}}$. We create occupancy matrices for each view and point set, i.e.,

$$O_j^{\text{back}}[u,v,w] = \begin{cases} 1 \text{ if any } \mathbf{p} \in \tilde{P}_j^{\text{back}} \text{ is} \in c(u,v,w) \\ 0 \text{ otherwise,} \end{cases} \tag{3}$$

and

$$O_j^{\text{forward}}[u,v,w] = \begin{cases} 1 \text{ if any } \mathbf{p} \in \tilde{P}_j^{\text{forward}} \text{ is} \in c(u,v,w) \\ 0 \text{ otherwise,} \end{cases} \tag{4}$$

where $c(u,v,w)$ is a unit cube at (ud, vd, wd), d is the length of the cube, and u, v, and w are the integer indexes of the matrices. From these, we compute the accumulation matrix

$$A[u,v,w] = \sum_j O_j^{\text{back}}[u,v,w] - \sum_j O_j^{\text{forward}}[u,v,w]. \tag{5}$$

Smoothing with a Gaussian function with width $\sigma = 1$ cm allows filling holes in the data, providing $A^*[u,v,w]$. Erroneous data points can, at least partly, be averaged out by using a sufficient number of views.

Before carving the object volume, we eliminate all entries in the accumulation matrix that are located behind the base plane of the turn table. This plane can be estimated from the markers (that have been previously used for finding the transformation matrices in Section 2.2). Therefore, let **a**, **b**, and **c** the 3D positions of the three markers that are located on the turn table, then the distance of a point (x,y,z) to the base plane is

$$\delta(x,y,z) = n_1 x + n_2 y + n_3 z - \alpha \quad , \tag{6}$$

where $(n_1,n_2,n_3) = (\mathbf{b}-\mathbf{a}) \times (\mathbf{c}-\mathbf{a})$ is the surface normal vector of the base plane, and $\alpha = \mathbf{n} \cdot \mathbf{a}$ a scalar distance. Then, with $x = ud$, $y = vd$, and $z = wd$, we set

$$A^*[u,v,w] = \begin{cases} A^*[u,v,w] \text{ if } \delta(x,y,z) < 0 \\ 0 \qquad\qquad \text{otherwise.} \end{cases} \tag{7}$$

Finally, we carve the volume by applying a threshold τ to the accumulation matrix and summing the non-zero elements, and obtain

$$V_{\text{est}} = \sum_{u,v,w} \theta(A^*[u,v,w],\tau) \quad , \tag{8}$$

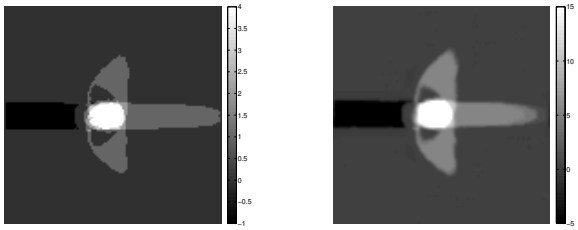
where $\theta(a,\tau) = 1$ if $a > \tau$ and zero otherwise.

3. Results

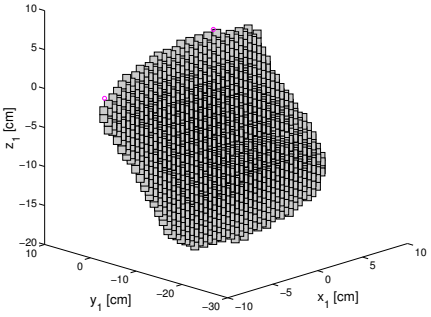
We applied the procedure to a total number of 16 household objects, including carton boxes, books, and cylindrical objects. In two examples, objects have been stacked on top of each other to obtain more complicated shapes. For all experiments, we used the same parameters, i.e., $\tau = 11$ and $d = 1$ cm. Color pictures of some of the objects are shown in Fig. 2. We measured the objects manually with a ruler in order to calculate the ground-truth volume V_{gt}.

We illustrate the main steps of the algorithm on the example of a paint can (see Fig. 2), a cylindrical object. In Figs. 3(a)-3(d), the point clouds for the four different views after background subtraction are shown in the camera coordinate system of the respective view. The point clouds can be merged by transforming them to a common coordinate system, i.e., the coordinate system of the first view, as shown in Fig. 3(e). This demonstrates that the cylindrical shape of the paint can has been correctly captured by the depth camera. We further observe that a large overlap between the different views exists, and alignment problems are notable in a few places.

As explained in Section 2.3 and 2.4, back- and forward projection of the acquired data points are performed in the respective coordinate system of the given view. The accumulation matrix is then computed according to Eqs. 3, 4, and 5. A slice of the color-coded accumulation matrix along the z_1 axis is shown in Fig. 4(a). Values accumulate in the area of the object (bright region) as compared to the area outside of the objects (darker regions). Smoothing with a Gaussian helps filling holes in the data (see Fig. 4(b)). Areas lying behind the base plane (defined by the plane of the turn table) are removed.

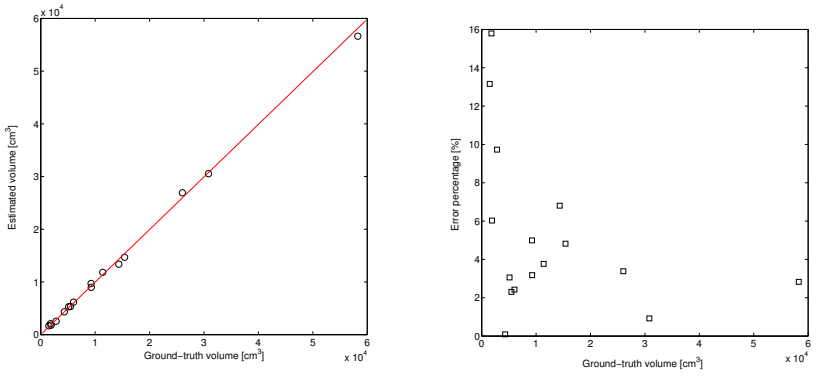


(a) 2D slice of A (b) 2D slice of A^*



(c) Volume carving

Figure 4. Accumulation matrices and volume carving: (a) 2D slice through the accumulation matrix $A[u,v,w]$. (b) 2D slice through the smoothed accumulation matrix $A^*[u,v,w]$. Bright areas correspond to the area occupied by the object. (c) Object volume approximated by unit cubes in the 3D space.



(a) (b)

Figure 5. Volume estimation results: (a) Estimated object volumes as a function of ground-truth volumes (squares). Object volumes could be approximated with an average error percentage of 5.2 %. (b) Error percentages as a function of the ground-truth volume. For very small objects, the error percentage is largest.

Table 1. Ground truth and estimated volumes in cm^3 of various household objects and error percentage

Object	V_{gt} in cm^3	V_{est} in cm^3	E in %
Cubic box	11390	11820	3.8
Shoe box	9282	8987	3.2
Paint can	4324	4320	0.1
Flour can	2816	2542	9.7
Beefeater box	1895	1781	6.0
House box	1792	2075	15.8
Nivea box	1482	1677	13.2
Shoe Box 2	5130	5287	3.1
Three stacked books	6020	6166	2.4
White box	5460	5334	2.3
Archivador box	9250	9712	5.0
Barcelona box	26040	26921	3.4
Robotis box	58275	56627	2.8
Unipro box	15410	14667	4.8
Two boxes unipro	30820	30535	0.9
Pccomponente box	14341	13365	6.8

By thresholding, the area of the object can be extracted. The extracted unit cubes of the object are shown in Fig. 4(c), representing the carved volume.

The volume of the object can then be calculated by summing up all the unit cubes according to Eq. 8. We obtained a volume of 4320 cm^3. In this particular example, a very small error of only 4 cm^3 compared to ground truth is observed, however, in general, for other objects, we observed errors to be in range of 4 to 1649 cm^3. Error percentages, defined as $100|V_{est} - V_{gt}|/V_{gt}$, ranged from 0.1 to 15.79 % with an average error percentage of 5.2 % (see also Table 1).

The differences with ground-truth values can be mostly attributed to (i) data discretization (unit cubes), (ii) data noise, and (iii) merging errors. In Fig. 5(a), the estimated volumes are plotted as a function of the ground-truth volumes for the different objects, following closely a linear 1-to-1 relationship, which demonstrates that object volumes could be correctly estimated within the observed error margin using the proposed procedure. We further show the error percentage as a function of the ground-truth volumes in Fig. 5(b). The error percentages are largest for smaller objects, which is due to the limited resolution of the procedure, affecting smaller objects proportionally more than larger ones.

4. Discussion

We investigated the use of consumer depth cameras for measuring volumes of household objects of arbitrary shape. Depth images were acquired with the Kinect for four different views of the object. Using a volume-intersection approach, the volumes of the objects were carved and compared to ground truth. To cope with faulty data near the object boundary, forward projection was included, counting negatively in the accumulation matrix for volume carving. An average error percentage of 5.2 % was found. In all ex-

periments the same parameters were used, demonstrating the robustness of the approach, given a fixed set-up.

To date, volumes of similar objects (parcels etc.) are still measured manually with a ruler in many places. Assuming a measurement error between 1 and 5 mm for each length (due to human error and irregular shapes or distortion of objects), we expect an error percentage for manual measurements between 1.6 and 7.9 % (calculated for the shoe-box example), which is in a similar range as the error of the proposed method. For a commercial volume measurement system (VMS420/520 from the company SICK), a length error ± 5 mm for similarly sized objects is given on the product sheet (see `http://www.sick.com/`), corresponding to an error percentage for the shoe-box of 7.9%, which is in a similar range as our approach. Hence, in a small warehouse scenario, consumer depth cameras as the Kinect could be used to measure objects with similar performance.

However, due to the limited resolution of the procedure, the approach is less suited for industrial tasks requiring high-precision measurements of very small objects. The precision of the method could be improved by using more views, improving the merging procedure itself, using smaller unit cubes, or by altering the sensor baseline and depth of field [15]. In a stationary set-up, multiple consumer depth cameras could be used to obtain different views of the objects instead of using a turn-table. Interferences between the cameras could be prevented by mounting a small vibrating motor to the cameras [16]. So far, we only measured the volumes of objects having fairly simple shapes. In the future, we aim to apply the proposed procedure to objects with more complex shapes, e.g, plants, vegetables and fruits.

Acknowledgements

This work received support from the CSIC project MVOD under project no. 201250E028. B.D. thanks the Spanish Ministry for Science and Innovation for support through a Ramon y Cajal program.

References

[1] G.P. Moreda, J. Ortiz-Cañavate, F.J. García-Ramos, and M. Ruiz-Altisent. Non-destructive technologies for fruit and vegetable size determination a review. *Journal of Food Engineering*, 92(2):119 – 136, 2009.

[2] M. Nylinder, T. Kubenka, and M. Hultnas. Roundwood measurement of truck loads by laser scanning. *Field study at Arauco pulp mill Nueva Aldea*, pages 1–9, 2008.

[3] F. Blais. Review of 20 years of range sensor development. *Journal of Electronic Imaging*, 13:231240, 2004.

[4] J. Battle, J. Mouaddib, and J. Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem: A survey. *Pattern Recognition*, 31:963982, 1998.

[5] P.J. Besl. *Chapter 1: Active optical range imaging sensors*. Springer-Verlag Inc., New York, NY, 1989.

[6] K. Hatou, T. Morimoto, J. De Jager, and Y. Hashimoto. Measurement and recognition of 3D body in intelligent plant factory. *Abstracts of the International Conference on Agricultural Engineering (AgEng)*, 2:861862, 1996.

[7] N. Sakai and S. Yonekawa. Three-dimensional image analysis of the shape of soybean seed. *Journal of Food Engineering*, 15:221234, 1992.

[8] D.J. Lee, J. Eifert, P. Zhan, and B. Westhover. Fast surface approximation for volume and surface area measurements using distance transform. *Opt. Eng.*, 42(10):2947–2955, 2003.

[9] K. Khoshelham and S.O. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 12:1437–1454, 2012.

[10] C. Loconsole, N. Barbosa, A. Frisoli, and V. Costa Orvalho. A new marker-less 3D Kinect-based system for facial anthropometric measurements. In *Proceedings of the 7th international conference on Articulated Motion and Deformable Objects*, AMDO'12, pages 124–133, 2012.

[11] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 559–568, 2011.

[12] A. Laurentini. How far 3D shapes can be understood from 2D silhouettes. *IEEE Transactions on PAMI*, 17:188–195, 1995.

[13] J.C. Carr, W. Fright, A.H. Gee, R.W. Prager, and K.J. Dalton. 3D shape reconstruction using volume intersection techniques. In *Computer Vision, 1998. Sixth International Conference on*, pages 1095–1100, 1998.

[14] J. Zheng. Acquiring 3D models from sequences of contours. *IEEE Transactions on PAMI*, 16:163–177, 1994.

[15] M. Ruther, M. Lenz, and H. Bischof. μNect: On using a gaming RGBD camera in micro-metrology applications. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 52–59, 2011.

[16] D. A. Butler, S. Izadi, O. Hilliges, D. Molyneaux, S. Hodges, and D. Kim. Shake'n'sense: Reducing interference for overlapping structured light depth cameras. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 1933–1936, 2012.