Citation: P. Artaso, G. López-Nicolás. Volume estimation of merchandise using multiple range cameras. Measurement, vol. 89, pp. 223-238, 2016.

Volume Estimation of Merchandise Using Multiple Range Cameras

Pablo Artaso[a], Gonzalo López-Nicolás[b]

[a] *Technological Institute of Aragón, C/ María de Luna 7-8, E-50018 Zaragoza, Spain*
[b] *Instituto de Investigación en Ingeniería de Aragón. Universidad de Zaragoza, C/ María de Luna 1, E-50018 Zaragoza, Spain*

Abstract

The ability of fast and automatic volume measurement of merchandise is of paramount importance in logistics. In this paper, we address the problem of volume estimation of goods stacked on pallets and transported in pallet trucks. Practical requirements of this industrial application are that the load of the moving pallet truck has to be measured in real-time, and that the measurement system should be non-invasive and non-contact, as well as robust and accurate.The main contribution of this paper is the design of simple, flexible, fast and robust algorithms for volume estimation. A significant feature of these algorithms is that they can be used in industrial environments and that they perform properly even when they use the information provided by different range devices working simultaneously. In addition, we propose a novel perception system for volume measurement consisting of a heterogeneous set of range sensors based on different technologies, such as time of flight and structured light, working simultaneously. Another key point of our proposal is the investigation of the performance of these sensors in terms of precision and accuracy under a diverse set of conditions. We also analyse their interferences and performance when they operate at the same time. Then, the analysis of this study is used to determine the final configuration of the cameras for the perception system. Real experiments proof the performance and reliability of the approach and demonstrate its validity for the industrial application considered.

Keywords: Range sensor, structured light, time of flight, volume estimation

Email addresses: artaso.pablo@gmail.com (Pablo Artaso), gonlopez@unizar.es (Gonzalo López-Nicolás)

1. Introduction

Volume estimation is a common problem in industry which is often time-consuming, complex and usually performed by human operators. Therefore, the automation of this task has attracted very often the interest of researchers. Despite related works have been proposed in the literature, their field of application is in general quite specific. Some examples are in the agricultural field for volume measurement of different kind of fruits [1] [2], or in the context of mining for estimating the in-bucket payload volume on a dragline excavator [3] [4], or in [5] for volume measurement of the load in a haul truck tray.

A common approach consists in measuring distances in the three principal axes of the target to compute its volume [6] due to its robustness and simplicity. In the field of medicine, some high-accuracy volumetric measurements have been designed using Robotic 3D Scanner. For instance, in [7] the *Frustum Sign Model* is explained to have an accuracy about 8% of measured volume, while the presented method reaches a 2% in controlled environments. However, it is difficult to find a system meeting the demanding industrial requirements and valid for a wide range of applications. More general volume measurement techniques are introduced in [8], which is specially appropriate for medium sized objects, or in [9], where dense surface reconstruction is pursued for volume estimation of objects with irregular shapes. A major contribution of this paper is the system presented which includes two new methods designed to measure volume in real time and in industrial conditions with an accuracy about 4,5%, which is quite acceptable for an industrial process that it has been performed manually until now. Up to our knowledge, there are not accurate generic algorithms for volume measurement such as the one proposed here. Therefore, we consider our proposal to have a great potential because our perception system is not constrained to a unique type of application, being applicable to a wide range of targets.

In all the works mentioned previously the volume is computed through depth information provided by range sensors. Currently, the major technologies to acquire depth information in computer vision are stereo cameras [10] [11], structured light [12] and time of flight [13] [14]. The stereo approach consists first in taking several images from different positions and viewpoints. Then, in order to get depth information it is necessary to match image infor-

Table 1: Main features of the devices employed in this work for data acquisition.

Sensor name	Asus Xtion Pro	Asus Xtion Pro Live	SR-4000	IFM O3D200
Manufacturer	Asus	Asus	Mesa imaging	ifm electronic
Technology	Structured light	Structured light	Time-of-flight	Time-of-flight
Range (m)	0.8 - 3.5	0.8 - 3.5	0.1 - 5	0.5 - 6.5
Resolution (pixels)	640 × 480	640 × 480	176 × 144	64 × 48
Field of view (°)	$58^o \times 45^o$	$58^o \times 45^o$	$43.6^o \times 34.6^o$	$40^o \times 30^o$
Frame rate (fps)	30	30	50	not specified
Output	x, y, z	x, y, z, colour (RGB)	x, y, z, intensity	x, y, z, intensity
Size (mm)	180 × 35 × 50	180 × 35 × 50	65 × 65 × 76	137 × 75 × 95
Weight (g)	230	230	510	1205
Power consumption (W)	< 2.5	< 2.5	12 (12V, ca. 1A)	16
Connection	USB 2.0	USB 2.0	Ethernet	Ethernet

Asus Xtion Pro: `http://www.asus.com/Multimedia/Xtion_PRO`
ASUS Xtion Pro Live: `http://www.asus.com/Multimedia/Xtion\_PRO\_LIVE`
Mesa SR 4000 `http://www.mesa-imaging.ch/products/sr4000`
IFM O3D200 `http://www.ifm.com/products/ind/ds/O3D200.htm`

mation and to know the intrinsic and extrinsic parameters of the cameras. Regarding the structured light technology, it basically consists in the analysis of the deformation of a known light pattern when it is projected into an object or scene. Then, the distance between the device and each part of the scene can be estimated by triangulation techniques. Finally, time of flight technology is based on the measurement of the time elapsed between the emission and subsequent arrival of a light beam, normally infrared, once it has been reflected into an object placed inside its field of view. Since its speed is known, the measurement of the range is immediate by computing the phase offset. As detailed in the following, the devices used in our system design are based on the two latter technologies.

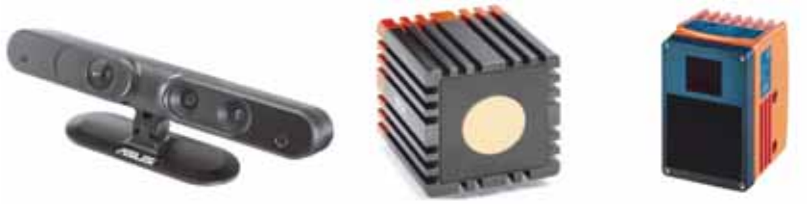


Figure 1: Devices used in this work. On the left an Asus Xtion Pro Live, a Mesa SR 4000 in the centre and an IFM O3D200 on the right.

In the last decade, range sensors based on structured light have been intensively developed and their considerable price reduction has eased their wide utilization in the research field [12] [15] [16]. Consequently, great efforts have been devoted to develop new techniques for processing and working with point cloud data. However, since these kind of sensors were initially designed as game controllers for domestic environments, they are not commonly used yet in the industry field due to their lack of robustness and protection required for industrial environments. On the other hand, the use of industrial oriented devices such as Mesa SR 4000 based on time-of-flight technology is becoming popular (e.g. [17] or [18]). The devices we consider in this work are Asus Xtion Pro, ASUS Xtion Pro Live, Mesa SR 4000 and IFM O3D200 (see Fig. 1). The two first are based on structured light whereas the others are based on time of flight technology. The data coming from these devices present important differences due to the distinct technologies they use to get the information which increases the difficulty of working with the data acquired considerably. Their main specifications are summarized in Table 1. A key issue when dealing with these kind of devices is the integration from the software point of view of the information they provide. A straightforward way to process this information is using the concept of point cloud because of its versatility to work with 3D data. In particular, what we use for image and point cloud processing in this work is the "Point Cloud Library" (PCL) [19], which contains numerous state of the art algorithms for 3D perception. Note that a key point that makes the problem addressed in this paper very challenging is the configuration of the cameras in the setup with very different points of view. State of the art approaches usually consider acquisitions of images with very close points of view to simplify the registration process. However, this can be a hard constraint in practice as it requires the acquisition of many images all around the target, which is not feasible for some applications. The method presented here only requires three images that can possess very different points of view to cover the merchandise. Standard registration algorithms will fail in fusing this information whereas our proposal handles this issue efficiently.

The aim of the perception system presented in this paper is to provide volume information through computer vision techniques during usual logistics processes such as truck loading and unloading or material flow controlling inside warehouses. This kind of automatic system for volume measurement allows the consequent time and cost saving. For addressing this task it is necessary to use multiple cameras simultaneously because, on the one hand, the information provided by a single camera device is not enough for registering the whole scene due to the dimensions of logistics merchandise, and on the other hand, it is not feasible in the application considered to rotate or manipulate the load in front of a single sensor to obtain complete information of the load. Next, several examples of multi-camera approaches that can be found in the literature are commented. A system consisting in a combination of ToF measurements with stereo in a semi-global matching framework is presented in [20]. Using several cameras allows increasing the working volume but it is also required extrinsic calibration to register all captured data. This issue is considered in [21] with ToF cameras. Another solution for multi-cameras systems is proposed in [22] by using a client-server-based capture system. Then, all captured data are transformed into point clouds using PCL, being afterwards all points fused into a global registered point cloud. However, in both previous works only devices of the same technology are used together, whereas the approach presented in this paper is able to combine information from very different devices such as Asus' cameras or MESA SR 4000. Although in terms of a real application this configuration would not be probably the first choice, we think that there can be applications with particular hardware constraints requiring such configurations. Then, the flexibility offered by our algorithms in comparison with the related works to perform correctly with so different devices is a clear advantage.

An important issue when dealing with multiple cameras simultaneously is the problem of interferences and the consequent performance degradation. This issue has to be considered in order to avoid jeopardizing the desired robustness and accuracy of the system. Works in the literature studying different factors related with these issues have been presented. Some examples are [23], which explains an exhaustive method for the comparison of the accuracy between the MESA SR 4000, Fotonic B70, and the Microsoft Kinect; [24], where a study of the effects of interferences on the accuracy and precision of RGB-D sensors depending on how they are placed is presented; or [25], in which a comparison study between Microsoft Kinect and Asus Xtion sensors is exposed.

In this paper, we present a different analysis to study extensively the performance of Asus Xtion, Mesa SR 4000, and IFM O3D200. Notice that, although [24] focuses on the study of the accuracy deterioration due to interferences, we are also interested about the response of the devices both in terms of accuracy and precision. Additionally, we compare and analyse the different devices considered (Table 1) in the same conditions and with materials typically involved in the considered field of application. In particular, we have focused on how their performances vary in terms of precision

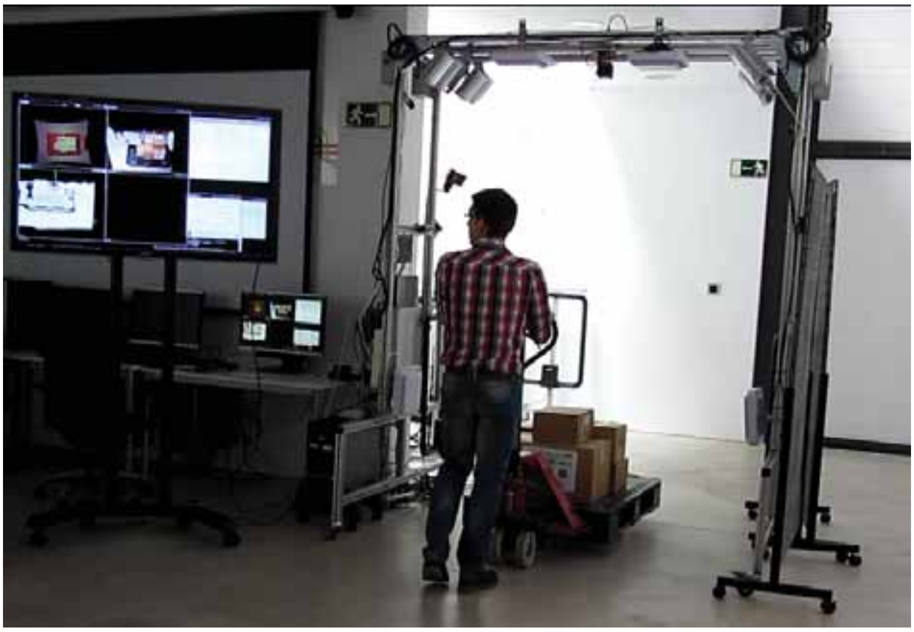


Figure 2: Environment where the system has been assessed. The three cameras used for volume estimation are located in the arch. The operator crosses under the arch pushing the pallet truck with the merchandise. The screen placed on the left side shows the user interface of the system and displays information in real time of the acquired data and the volume estimation.

and accuracy when working conditions such as distance or target materials change. Additionally, it has been studied how they are affected when they operate with another device at the same time. The conclusions of this study have been useful for locating the sensors in an adequate configuration for the task addressed.

Thus, the first contribution of this work is the analysis and comparison study of the cameras aforementioned. The second contribution is the design of two robust, flexible and accurate algorithms for volume estimation starting from the data acquired by quite different cameras. Then, these algorithms have been implemented in a perception system to automatize the procedure. The system setup to evaluate the proposed volume estimation algorithms consists in a metallic arch where the cameras have been installed. An image showing this setup with an operator pushing a pallet truck while the system is working is shown in Fig. 2. The proposed system has been designed to estimate the volume of the load stacked on a pallet when an operator transporting it employing a pallet truck passes through the arch, although it can be used in other scenarios given its flexibility. Then, a screen next to the arch shows to the user the volume estimation in real time. It must be remarked that due to the flexibility of the designed algorithms the system can be easily adapted to other transportation devices too or even to other scenarios.

The paper is organized as follows. In Section 2 is presented an analysis of the performance of multiple range sensors and the interferences when they operate at the same time. The following Section 3 describes the segmenting pipeline to extract the useful information from the point clouds acquired. In Section 4, 3D registration process is explained. The algorithms designed for volume estimation are described in Section 5. The proposed approach is tested with fifteen real situations in Section 6. Finally, conclusions are given in Section 7.

2. Analysis and comparison of cameras

Before addressing the volume computation process, we found it essential to gain further technical knowledge about the performance of the devices. In particular, it is important because it is very difficult to develop an accurate application for volume measurement whether the measuring depth average error is high. Therefore, both precision and accuracy of the devices have been studied in different working conditions. The trials considered different working materials, as well as their distance to the devices. Then, it is possible to compare the devices and the different technologies involved. Additionally, another purpose of the study is to know whether interferences exist when several devices work together, and in that case, which are their effects. Besides, the study of the precision and accuracy has been used to better define the configuration of the devices in order to improve data quality and consequently, get better results with the system in the task of volume estimation.

A variety of materials have been considered in this study such as a cardboard box, a plastic box, a wooden pallet, a plastic pallet, several sheets of paper and cards of distinct colours, a bubble wrap plastic, a greenhouse plastic, a metallic bin or a laptop screen with wallpapers of different colours. Then, one hundred point clouds have been acquired for every object at two different ranges, 1.5 and 3.0 meters, for the purpose of analysing how data is affected. Fig. 3 shows the materials used in the study and how the devices have been placed.

Another important aspect we have studied is the influence of varying the illumination conditions. In particular, we considered typical indoors electric light and indirect sunlight. The comparison has not been included since it shows that the effect in the performance of the cameras under normal operational conditions is negligible.

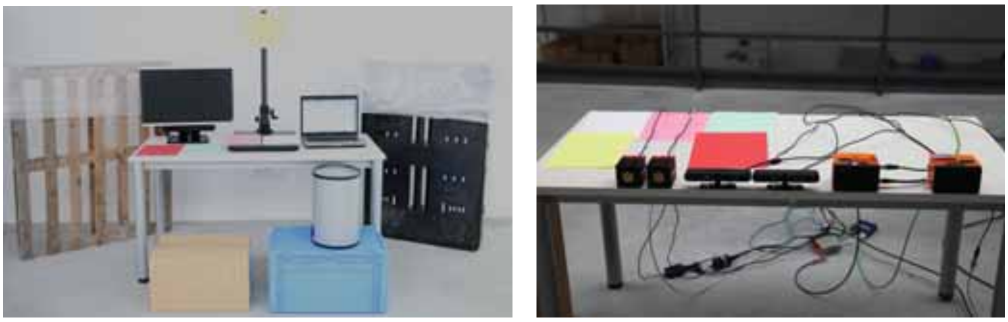


Figure 3: On the left, the elements employed in the study of precision and accuracy. On the right, the devices and their configuration in the study.

2.1. Precision and accuracy study

Firstly, it is convenient to highlight the properties we are going to study in order to analyse the data adequately. On the one hand, precision is the closeness of agreement between indications or measured quantity values obtained by replicating measurements on the same or similar objects under specified conditions (according to the International Vocabulary of Metrology (VIM)[26]). On the other hand, accuracy is the closeness of agreement between a measured quantity value and a true quantity value of a measurand. Precision has been estimated through the standard deviation of the range measured by the sensors, which represents their variability; whereas accuracy has been estimated through the error between the correspondent value of the z-axis (orthogonal to the camera image plane) and the real distance to working materials.

Precision results for both, 1.5 and 3.0 meters, are shown in Fig. 4. Attending to Asus, variability considerably increases in all the situations when distance raises (an increment of 250% on average). However, this effect is not so pronounced for IFM or Mesa. In fact, Mesa's precision improves for some materials, specially those made of plastic. Therefore, a first conclusion is that although the precision of these devices was initially slightly lower than Asus,

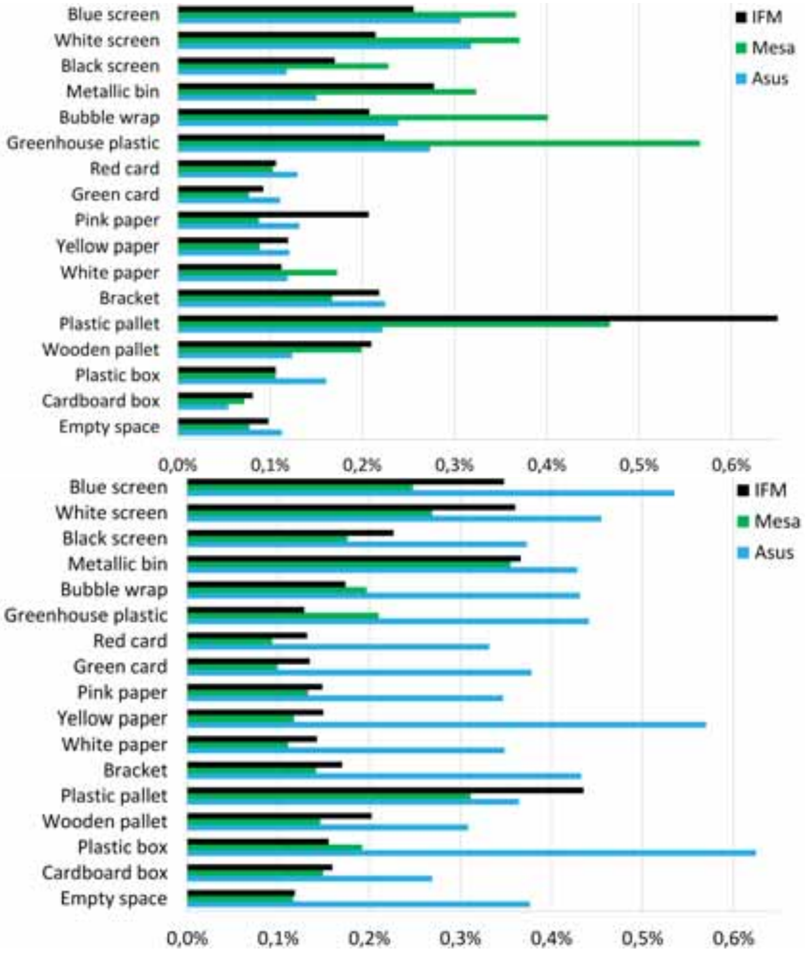


Figure 4: Both graphics show the mean of the standard deviation of the range computed in each pixel of the devices for all the materials used in the study. They represent the variability of every range sensor in each pixel, what it is to say, their precisions. The first graphic displays the results when the working distance is 1.5 meters and the last one when it is 3.0 meters.

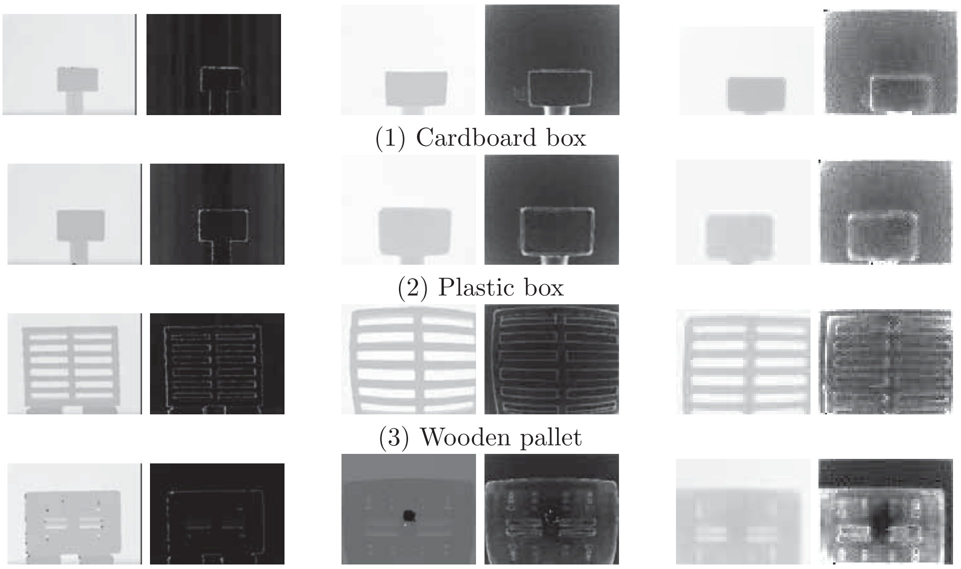
it remains quite stable when the distance shifts. The major reason to such behaviour is the distinct aims these devices have been devised to, because whilst Asus' cameras are game controllers, Mesa and IFM are industrial oriented sensors and therefore, designed under quite different constraints. More detailed explanations are exposed in the presentation of the accuracy study results.

Another remarkable aspect in Fig. 4 is that both Mesa's and IFM's devices present bigger variability differences when working materials change, being much more pronounced when the materials are composed of plastic. This is specially striking in the case of working with the plastic pallet, where the precision of both devices is pretty bad. The main reason to this performance is that the pallet has been chemically and physically processed, resulting then in a smooth surface with a high degree of specularity, so it reflects too much light, what deteriorates the measurements based on the time of flight technology.

In conclusion, it can be stated that the Asus presents a more stable performance generally, or, in other words, it is less affected by the working objects than the other devices. The principal reason to this performance is the different technologies the devices are based on. Since time of flight depends on the intensity received from the beam reflected into the materials, if the material changes, the same does the quality of the data received. However, structured light does not depend so much on the properties of the material, so it is considerably less affected by this issue. Note also that the colour change does not affect to the performance of the devices, according to the similar results obtained with the different papers, cards or even the computers' screens.

Given the necessity of knowing the standard deviation in each pixel of the devices to analyse their precision, this information can be used to represent how the variability is distributed throughout all the scene for each camera. This means that it can be known which parts are the most problematic for every range sensor. For this reason, one hundred point clouds of each scenario have been acquired to measure the precision in each pixel. Then, depth data of all the point clouds have been analysed for each pixel to compute their standard deviations. That is to say, we measure the variation of depth information in each pixel during the acquisition of one hundred point clouds. A greater variation is linked with a less reliable data, which usually means a higher error in the volume measurement process. The parts of the scenarios with higher variability have been represented with white colours in the second row of pictures for each sensor in Fig. 5.

In the case of Asus, it can be observed in the second column that the higher variability basically appears throughout the contours. Even though the same occurs with the Mesa, it does in a less intense way. However, it can be checked that the variability of the inner data of the objects is a bit higher, mainly when they are made of plastic. From the images of the



(1) Cardboard box

(2) Plastic box

(3) Wooden pallet

(4) Plastic pallet

Figure 5: The two first columns correspond with the images acquired by Asus Xtion Pro, the next two by Mesa SR 4000 and the last by IFM O3D200. In the first column of each pair is shown the depth images for the different objects used: a cardboard box, a plastic box, a wooden pallet and a plastic pallet respectively from top to bottom. Depth estimation variability can be observed in the second column for each point of the scene, in a scale from black to white (the biggest variability).

IFM, it is noteworthy that there is barely difference between the variability of the contour and inside the object. Additionally, the variability is rather distributed in all the object. There are reasons which explain the lower precision in the contours depending on the technology used to get the depth information. In the case of structured light devices, there are shadows in the triangulation process done to compute depth as a consequence of being the contour zone visible only for one of the two artefacts which form the device (light emitter and light sensor). In the case of time of light technology, there are multiple reflections which falsify data when the beam reaches the edge since the surface is not totally plain.

Next, it has been studied the accuracy of the measurements in the z-axis. Specifically, we have computed the accuracy through the mean of the average values measured in each pixel of the devices for one hundred point clouds acquired for each distinct object. The results are shown in Fig. 6 for both working distances, 1.5 and 3.0 meters. As with the precision, the Asus

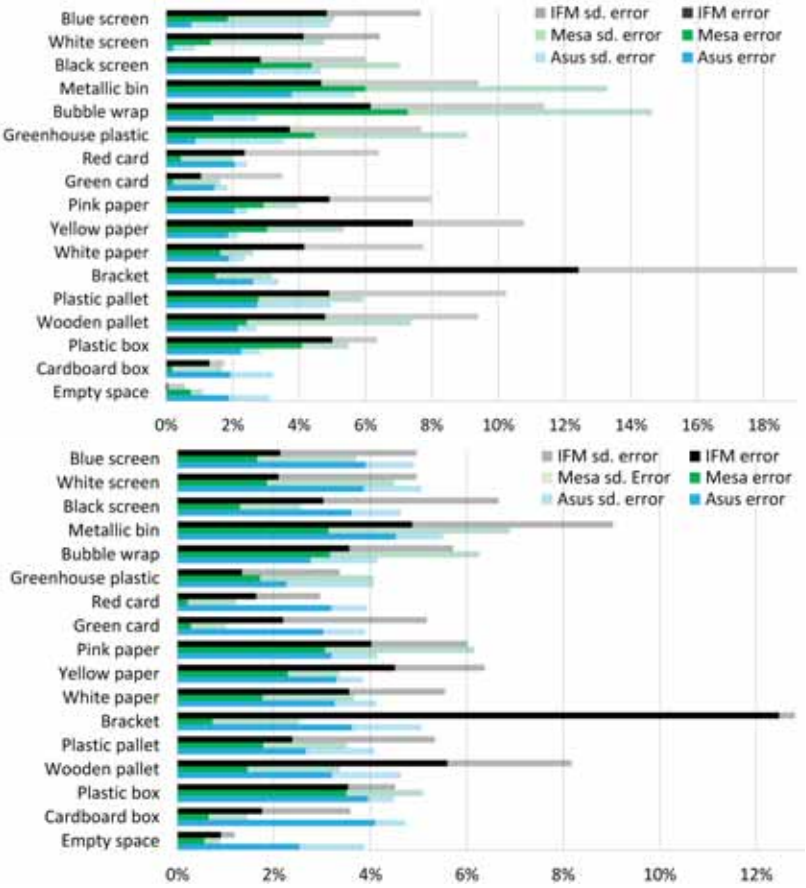


Figure 6: Mean error when measuring range with the cameras for the different objects and materials at 1.5 meters and 3.0 meters (first and second graphic, respectively), which represents the accuracy of the sensors.

presents the best accuracy for the working distance of 1.5 meters, being fewer than 3%. However, it considerably deteriorates when the range increases to 3.0 meters. Even so, this increment is not as big as in the precision study. In addition, the results are quite homogeneous attending both precision and accuracy. Hence it can be stated that Asus is very robust faced with working material. Another important aspect is that the standard deviation of the accuracy is smaller with the Asus than with the other devices in all the situations by far.

Analysing Mesa, it can be observed it has better accuracy than the IFM in almost all the cases, being the worst response with the materials composed of plastic, as it happened in the precision study. However, both sensors improve their performance when range increases, specially in the case of Mesa. One of the reasons to this performance behaviour when the object is further away is that time of flight technology is basically based on measuring the time an emitted wave takes to come back to the sensor. Nowadays, it is relatively easy to measure this time elapsed with great accuracy, being even easier when this time is higher. Another reason is the different aims these devices have been devised to. Asus' cameras have been initially designed as game controllers, so their application do not require a great quality on the acquired data. However, IFM and Mesa are industrial oriented cameras, so they are used in applications where accuracy and precision can be essential.

Although in the viewpoint of the precision there is no difference between working with papers or cards, on the other hand it does in the accuracy of time of flight-based cameras. The reason is that paper is thinner, therefore the range measurements are distorted because the infrared beams go partially through it.

2.2. Interference Study

Regarding the performance of the devices, it is not only necessary to know their precision and accuracy features, previously analysed, but also it is important to know about their operation working with other devices simultaneously. This is capital for the correct performance of the proposed system, which is devised to work with several cameras in order to make a 3D registration from the information of the merchandise acquired from different viewpoints in one shot.

For evaluating wether interferences appears between the devices, the variations of the precision and the accuracy have been studied for each device working with others together, also including the case in which two equal devices work simultaneously. The tests have been carried out for 1.5 meters and employing some materials of the precision and accuracy study such as a card, a paper, a cardboard box or with empty space to check whether the change of the working material influences in the interferences.

The results of the interference study are depicted in Figs. 7, 8, and 9 corresponding to Asus Xtion Pro, Mesa, and IFM, respectively. From the collected data, it can be concluded that the single situation in which Asus suffers from interferences is working with another Asus. The reason is that both patterns mix, then their correct performance is interfered and as a result

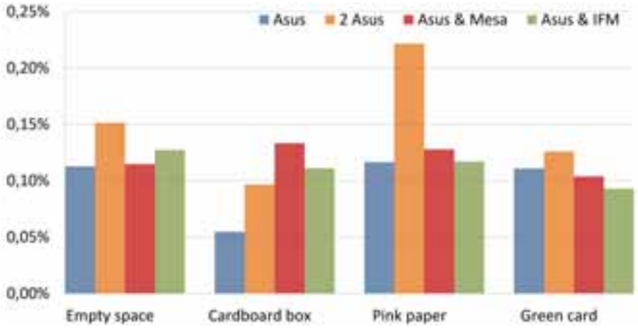


Figure 7: Mean of the standard deviation of depth measurement in each pixel of the Asus when the working distance is 1.5 meters and other devices are operating at the same time.

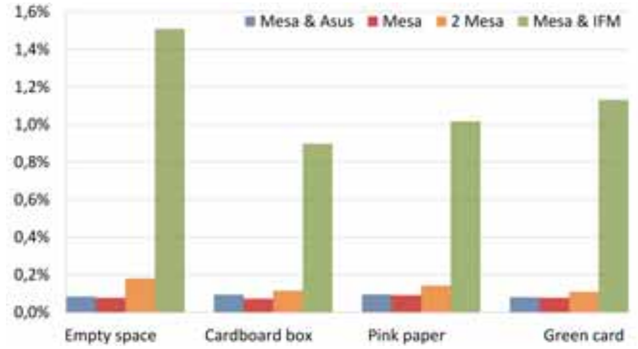


Figure 8: Mean of the standard deviation of depth measurement in each pixel of the MESA when the working distance is 1.5 meters and other devices are operating at the same time.

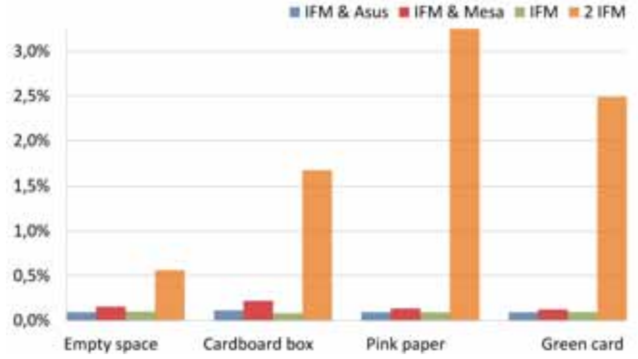


Figure 9: Mean of the standard deviation of depth measurement in each pixel of the IFM when the working distance is 1.5 meters and other devices are operating at the same time.

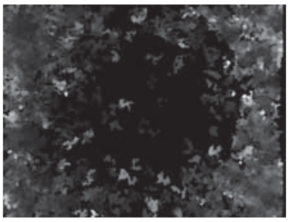


Figure 10: The image represents the variability in each pixel of an Asus (where the lighter colours represent higher standard deviations of the depth measurement) when its IR emitter has been blocked and a second Asus is placed next to it. The cameras are pointing the same point of a wall in this example.

their variability increase. Despite the previous issue, the effects of the interference do not seem severe because the accuracy is not affected and in general there is a slight increment of the variability, with the exception of the paper where it is a bit higher. For instance, in Fig. 10 it is shown how the pattern of a second Asus affects to the variability of the first Asus. Recommendations to place structured light cameras in order to minimize interferences like Fig. 10 have been proposed in [24]. It is important to say that even though the results shown in Fig. 7 demonstrate there is not interference between the IFM and the Asus from the point of view of precision, it has been visually checked the apparition of holes into the Asus' point clouds when they work together (Fig. 11). It has been confirmed that the appearance of this phenomenon depends on the IFM's exposure time, which should be modified depending on different aspects as the working material, the lighting or the range.

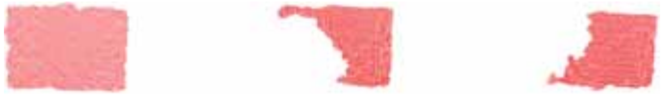


Figure 11: On the left, an Asus' point cloud of a card. In the other two images, the same card when the Asus works with an IFM simultaneously. The holes in the two latter are a consequence of interferences caused by IFM.

Attending to Mesa's data (Fig. 8), a deep interference when it is working with the IFM simultaneously is observed. Besides, there is a slight increment of the variability when two Mesa work together. It must be highlighted that both Mesa are synchronized, otherwise results would be similar to the case of one Mesa working with an IFM. On the contrary, by analysing the IFM interferences (Fig. 9), it is surprising that there is few influence working with the Mesa simultaneously. In this case, the increment of the variability is minimum (around 0,1%) compared with the correspondent to the Mesa (around 0,8% or more, depending on the material). In view of Fig. 9, the biggest interferences occurs when both IFM work at the same time, causing not only a high variability increment but also an accuracy decrease working with certain materials. The main reason for this performance is that they have not been synchronized.

2.3. Analysis and discussion

Once the precision, accuracy and interferences of the devices have been analysed, some conclusions have been reached for the design of the system. First, an IFM working with other devices can be counterproductive. Even though the detected problems between an IFM O3D200 and a Mesa SR 4000 can be solved with a correct synchronization, it is not possible to solve the observed interferences when an Asus' camera works with an IFM simultaneously. This is one of the reasons why we have decided not to use the IFM's devices in the volume estimation system. Another reason is that the accuracy of Asus is better than IFM for short distances. Besides, the Mesa's error is lower than the error of IFM for long distances.

Moreover, there is one more reason for such decision. From the point of view of the resolution, the IFM's camera presents a clear disadvantage (Table 1). For example, in the case of working with the cardboard box placed at 1.5 meters, the number of points that represent such object with Asus are 20409, with Mesa 3130 and finally 272 with IFM. If the working distance is 3.0 meters, the number of points is 4795, 799 and 87 respectively. Then, the few points that represent de box in the case of the IFM can cause the loss of some details, which is crucial to achieve an accurate system. In particular, the side of the box is 50 cm and it is described by 22 points with the IFM at 1.5 meters, so for at least 2,3 cm there is not any information available, which can increased the volume estimation error obtained by the system up to 10% in the worst case (considering only the error in one axis). Therefore, the cameras finally employed in the proposed application are Mesa SR 4000, Asus Xtion Pro and Asus Xtion Pro Live.

2.4. Spatial configuration of devices

A critical point of the system is the placement of the devices. It determines some aspects such as their operation range, how the point clouds acquired are or how the information represented by them is. It is also necessary to take into account that the devices must have some intersecting view in order to be possible to make a 3D registration of the merchandise.

The conclusions of precision and accuracy study have conditioned decisions about how to set the cameras. In the study has been proved that working range affects in a different way to data quality of each device. Specifically, Asus shows a higher performance when the range is smaller. However, Mesa provides better data when range is bigger, according to how its precision and accuracy change. Additionally, taking into account that the considered application requires not blocking the pallet truck path, we chose to install the devices in an arch shape metal structure appropriate for using it in a corridor. For these reasons, we have decided to set two Asus at the structure sides and one Mesa at the top part. Since the structure sides are closer to the merchandise than the top, this setting favours that cameras work in optimal distance conditions. In addition, this configuration allows minimizing interferences between both Asus.

Although it has been decided to work with two Asus and one Mesa, other set of devices could have been chosen, however they present some drawbacks which are described below:

- **Three Asus**: the interferences where the three patterns intersect are substantial, which coincides in where the merchandise passes. As a consequence, the accuracy of the system would decrease.

- **Three Mesa**: these devices must be as faced as possible to the merchandise. So in case they are placed at the sides of the corridor, they should not be in a slanted orientation, but otherwise they would be too close to the merchandise. Consequently there would not be enough minimum distance to work correctly.

- **Two Mesa and one Asus**: if the Mesa's devices are placed at the sides, the problem is the same than in the latter situation. In case they are placed at the top of the structure, we would not dispose of enough information of the sides of the merchandise because only the Asus remains to be placed.

Note that, although additional cameras could be used to obtain better coverage of the merchandise, we choose only three in order to keep acquisition and processing data time bounded as well as to limit the system cost. Once

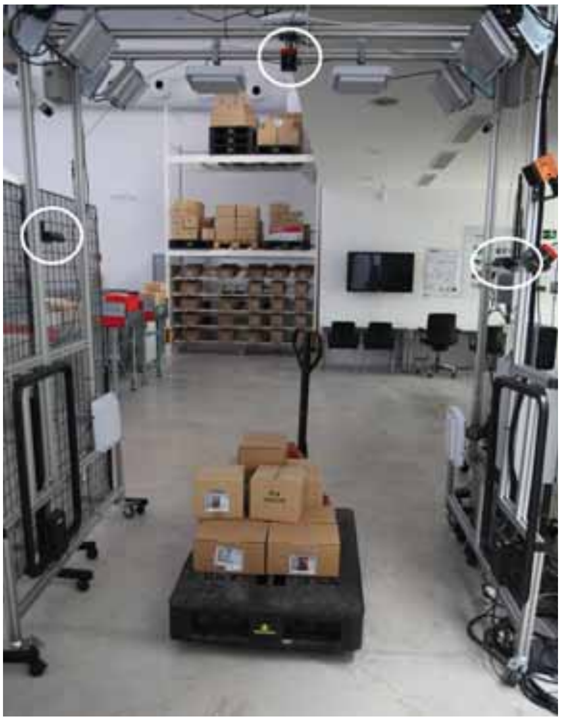


Figure 12: Configuration chosen for the devices. Both Asus on the sides and the Mesa on the top part. The Asus registers lateral information exclusively and the Mesa gives an upper view of the scene, being its z-axis perpendicular to the floor plane. The cameras have been highlighted in the image with white circles.

placement of devices has been determined and which of them are being used, there is still to decide their orientations to face the load. In the case of Mesa, we have decided to face it to the floor, that is to say that its z-axis matches up with the perpendicular to the ground plane. Such orientation eases some aspects of the algorithm because its point of view is a zenithal view where the z-axis matches up directly with the height of the objects. It must be remarked that the own manufacturer of Mesa recommends for a correct performance to face it perpendicularly to the target because otherwise malfunctions related to multiple reflections can happen (Fig. 13). Asus' cameras have been oriented to the sides of the merchandise, but slightly inclined to the ground. The final configuration of the three cameras can be seen in Fig. 12.

The only drawback of the chosen setting is that front and rear information

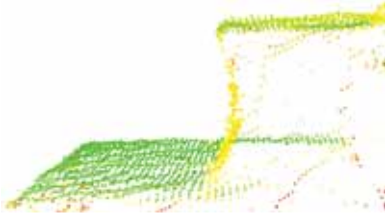


Figure 13: Note that the side of the box represented in the point cloud acquired by MESA SR 4000 seems curved when it should be plane. Besides, the majority of points which represent it are concentrated on the lower part. This phenomenon working with the Mesa is due to multiple reflections caused by the proximity of other planes to the curved plane. As a result, the beams rebound in these other planes, and the measurement is distorted. The device is placed in the top of the arch facing the front part of the merchandise (approximately on the left top of this picture).

of the merchandise we want to estimate its volume is not available because enough free space is needed for the movement of the merchandise. However, it is not a problem because side and top information of the target is enough for volume computation as we demonstrate in this work.

In particular, all cameras are oriented to the same area and they are triggered at the same time to synchronize data acquisition. Note that during the trials, information was acquired under request of the operator. It must be highlighted that the merchandise is supposed to be rigid, that is to say that there are neither deformation nor relative movement between the boxes. Therefore, any possible de-synchronization does not affect the results as long as the information across the sensors is properly matched, as it is proved in the experimental results.

3. Scene segmentation

First of all, it is necessary to unify the data from the devices under a single format. For this purpose, we have chosen to work with the Point Cloud Library (PCL) to manage point clouds, since it provides tools facilitating to deal with and process this data. PCL allows to control the Asus' cameras and it converts the recorded data into the desirable format too. In the case of Mesa, we implemented the code both to control the camera and to convert data to the required format.

Even though all point clouds we work with are under the same format, they are very different because of the disparate ways the devices they come from acquire the data. An example is the type of information provided by the devices. On the one hand, the Asus provides depth and colour information of the scene, so overlaying them a 3D colour point cloud is granted. But, on the other hand, the Mesa does not incorporate a RGB sensor so it is impossible to get colour information. In this case, Mesa provides depth and intensity information obtained by the difference of energy between the emitted infrared light beam and the received one after being reflected into the objects placed inside its view. Then, through the union of both data a 3D intensity point cloud is achieved.

As it has been already mentioned above, the aim of this paper is the volume estimation of the merchandise transported on a pallet. However, since there are more objects into the scene apart from the merchandise (the floor, a pallet truck, etc.), it is necessary to process the acquired data to work only with useful information through the removal of the corresponding data to the pallet truck, the ground or the pallet. The scene segmentation is a typical problem in computer vision and different approaches are usually used depending on how the scene is [27]. In our project, the processing pipeline consists of a plane extraction for the floor, the application of filters for the removal of the pallet, followed by a clustering for isolating the merchandise. The use of these well-known techniques that have been widely applied provides robustness to the system (e.g. [28]). In the following sections, we describe the procedure to extract the surrounding elements of the merchandise.

3.1. Ground extraction

So as to extract the ground, the widely known Random Sample Consensus algorithm (RANSAC)[29] has been used. However, instead of applying it every time a point cloud is acquired, we propose to calibrate the floor plane through RANSAC the first time cameras are placed in the workplace. Thereby, it can be granted that, during the calibration, there are only ground and walls inside the view of the cameras. Thus, the number of spurious points is substantially reduced during the estimation of the ground plane equation for the purpose of the floor being the main plane of the scene and therefore ensure the robustness of the ground extraction.

Afterwards, each time it is needed to extract the ground plane in order to do a new volume measurement during the system normal operation, it is solely necessary the removal of those points belonging to the plane previously computed. Proceeding in this way makes sense because once the cameras have been placed, the floor equation is always the same, so it would not be efficient to apply RANSAC each time the floor needs to be removed.

3.2. Pallet truck and pallet extraction

Even though the ground is easily identifiable, the extraction of the rest of components can be more complicated. Then, we have decided to use the additional information the cameras provide to extract the pallet from the merchandise, which are the intensity and the colour of the scene. This is possible due to the features of the employed pallet, which is composed of polymers. The usage of such type of pallets is increasing because they are recyclable, resulting in an important saving. In this way, features like colour are readily modifiable. In our case, the pallet used is black, and as a consequence it can be easily detected. Additionally, this kind of surface is in general smooth and with a high level of specularity. As a result, the intensity received from the pallet in the Mesa is considerably lower than the intensity from the rest of elements of the scene because it is highly scattered in every direction.

Therefore, by applying a colour filter we can remove almost all the pallet from the point clouds acquired by the Asus with colour sensor. In the case of Mesa, the procedure is analogous but over the intensity information. This method has been applied because the height of the pallet can be different each time given that the user can vary this height arbitrarily. Consequently, it cannot be calibrated offline and online method as the one proposed here is necessary. However, since Asus Xtion Pro does not integrate a colour sensor nor an intensity sensor, none of the previous filters can be applied. For this reason, we have estimated a mean height of the pallet with respect to the floor for the Asux Xtion Pro's filter. As the equation of this plane is known, the pallet can be roughly extracted by removing those points with a lower distance to the plane than this measurement.

At this point, it should be applied the filter explained in the following section 3.3 that we have developed. The purpose of this filter is solving reflections problems detected working with the Mesa SR 4000 and the plastic pallet in some specific situations, specially when the surface of the pallet where the Mesa just aims straight is not covered with any object.

Once the pallet has been extracted, there might be only a few secondary elements in the point cloud, apart from the load, such as part of the pallet truck or the operator. A common technique in 2D image processing called Euclidean Cluster Extraction has been used for extracting them, but applied

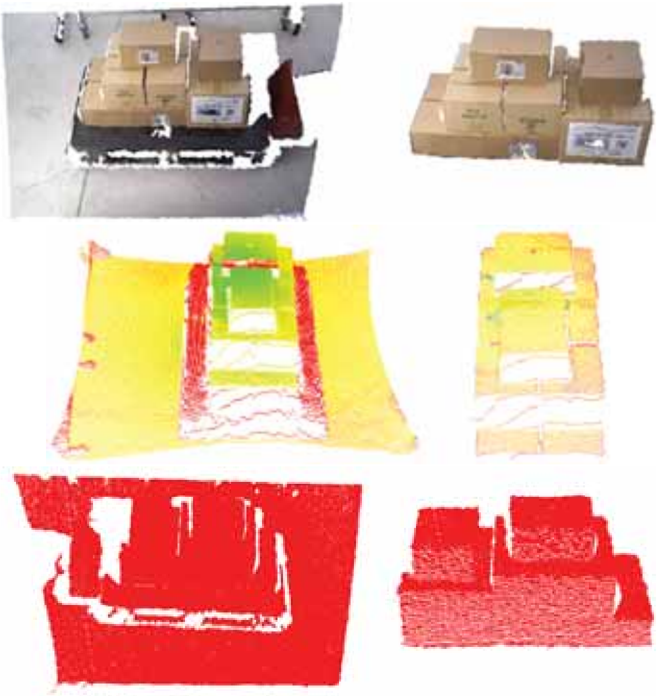


Figure 14: Processing example for a point cloud acquired by the Asus Xtion Pro Live, Mesa SR 4000 and Asus Xtion Pro respectively. The first picture of each pair is the point cloud acquired directly and the second the result of the segmentation process, where only the merchandise remains.

to 3D point clouds [30]. If more than one cluster is found, only the one with the highest number of points would be considered to be the merchandise. An example of point cloud processing for every device is shown in Fig. 14.

3.3. Filter of reflections

Throughout the trials, reflections in the point clouds acquired by Mesa due to the specular characteristics of the pallet surface were observed (Fig. 15). As a consequence, points belonging to the pallet are not removed in the scene segmentation process, deteriorating the merchandise volume estimation.

For this reason, a reflection filter for the Mesa has been designed starting from the two point clouds resulting of applying the intensity filter explained in section 3.2 to Mesa's point clouds. One of the point clouds is composed of those points with an intensity value over the filter threshold, which mostly belong to the merchandise and the undesired reflect (green and yellow points

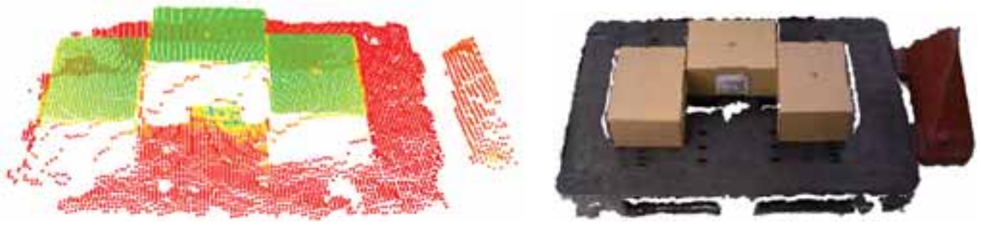


Figure 15: Instance of two point clouds where there are reflections. In both the floor has been extracted previously. On the left, it can be noticed a Mesa's point cloud with a reflection in the pallet (green points which should be red) caused by the own camera. On the right, a point cloud taken by the Asus Xtion Pro Live simultaneously which is not affected by this phenomenon.

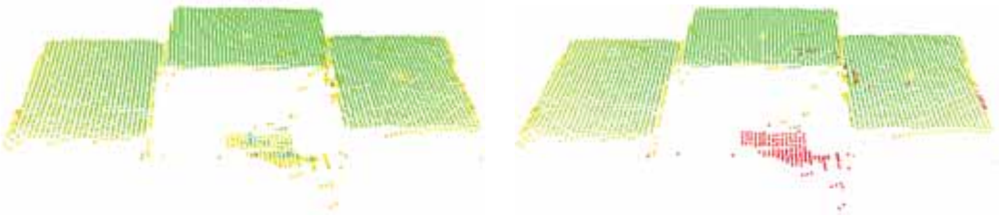


Figure 16: On the left, the point cloud of Fig. 15 after being segmented and before applying the filter of reflections. It can be noticed the reflection in the middle which belongs to the pallet (it should have been removed in the segmentation process). On the right, the result of applying the filter, where the reflection has been identified correctly (red points).

in Fig. 15). The other point cloud includes the rest of points, which correspond with the pallet (red points). The steps summarizing the process are explained in the algorithm 1. The result of applying this filter to a point cloud is shown in Fig. 16, in which can also be observed the differences between applying the filter of reflections or not to the same point cloud.

4. 3D registration

Once all the point clouds acquired by the devices have been processed, all the required information to estimate the volume of the merchandise is disposed. We also need to take into account that the reference system of each device is different, consequently it is necessary to transform the point clouds under a unique reference system, that is to say, do a 3D registration. The problem of merging the information amounts to find the transformation matrices in $\mathbb{R}^{4\times4}$ which contains the necessary information to rotate and

Algorithm 1: Summary of the Reflection Filter Process.

```
input : Two point clouds obtained through the application of an
        intensity filter into a Mesa's point cloud, the first depicting
        the merchandise (points with higher intensity) and the
        second the remaining elements (lower intensity).
output: Point cloud without the reflection.

Transform the point cloud with the lowest intensity points into a 2D
  image. Recall that in this image the points corresponding to the
  merchandise and the reflection are not included;
Apply a *dilating* and an *erosion* process for filling the smallest holes
  which virtually correspond with the reflection and some corners of
  the merchandise;
Compare point-to-point this image with the correspondent to the
  segmented merchandise. Those points present in both images are
  candidates to belong to the reflection. However, it is possible that
  some contour points may be classified as candidates too;
To remove the reflection from the outline of the merchandise, search
  for the set of points more centred and closer to the floor. In case a
  point belongs to the reflection, its height would be lower than the
  height of the rest because the pallet is the closest element to the
  floor. Besides, in that case it would be just located in the centre of
  the image because the reflect appears just underneath the device as it
  can be observed in Fig. 16;
```

translate the point clouds between different reference systems. As a result, they allow to change the reference system of a point cloud acquired from one device to another. This problem has been extensively studied [31] [32] [33] and the steps to carry out can differ from the type of point cloud used or how is the scene depicted.

In the following, we illustrate how to transform the reference system once the transformation matrix $\mathbf{T} \in \mathbb{R}^{\mathbf{4\times4}}$ is known.

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}, \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{\mathbf{3\times3}}$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^{\mathbf{3}}$ is the translation vector between the reference systems of two cameras. Given a 3D point $\mathbf{X_i^k}$ from a point cloud acquired by sensor k and expressed in its own reference with homogeneous coordinates, it can be transformed to a common reference of a second sensor h by applying the transformation matrix $\mathbf{T}$ which links both reference systems:

$$\mathbf{X_i^h} = \mathbf{T} \cdot \mathbf{X_i^k} \tag{2}$$

The solution proposed to compute the transformation matrix which assures a 3D registration of the merchandise without affecting the robustness of the system consists in doing a unique initial calibration of the devices. That is to say that the first time the system is installed, the user will select some keypoints from the scenario and then, the user will define correspondences from these keypoints between the different point clouds, each one acquired by a distinct camera. Note that it is not required any calibration pattern. In particular, it is only required a default scenario up to the user's choice. The only specification for this scenario is that it should have distinguishable points among all the target space where the volume will be measured and also these points must be observable in at least two different point clouds. From the correspondences, a first approximate transformation matrix is calculated through the application of the algorithm developed by Arun et al. [34] based on the singular value decomposition (SVD). The transformation matrix obtained by this way is employed as a seed for the Iterative Closest Point algorithm (ICP) [35]. Owing to the correctness of the correspondences introduced manually, the seed is quite close to the final solution which eases the convergence of the ICP to the minimum global error and in just a few iterations.

Proceeding in this way, we avoid doing all the typical procedure of an automatic 3D registration each time the volume is computed, resulting in a considerable save of computation time. Such approach is possible because relative positions are always the same once the cameras have been placed. Therefore, the transformation matrices that connect the different viewpoints of the three devices are constant. So once they have been calculated, it is only necessary their application to change the reference system of the point clouds initially acquired.

The automatic computation of the transformation matrix each time a point cloud is acquired is not used because of the high computation time and the lack of robustness observed in testing phase. The main reason is that only three cameras are being applied to register all the merchandise, so the change of viewpoint from one device to another is very significant and the baseline is very large too. Indeed, we are trying to register a 180^o arch with only three cameras. In addition, another reason which explains this performance is that the scene is normally composed of boxes. Therefore, almost all the geometric elements we are working with are rather similar because they consist of a set of parallel and perpendicular planes. Moreover, attending to the levels of intensity or colour registered by the cameras, they are considerably homogeneous. In conclusion, it is really tough to automatically obtain appropriate keypoints and good correspondences with the required robustness for an industrial application due to the large baseline, the lack of intense gradients or characteristic areas in the scene.

5. Volume estimation

In this section, we present the methods developed to estimate the volume of the merchandise. Without loss of generality, we have assumed for the design of the algorithms that there are not holes nor projecting elements inside the load due to the limitations of the sensors and for security reasons.

5.1. Voxel method

Algorithm 2: Volume estimation with *Voxel* method

```
input  : 3D registration point cloud and the equation of the ground plane.
output: Volume of the merchandise.
Align the $z$-axis of the point cloud with the normal of the ground plane;
Search the point cloud for the point with the minimum distance to the floor plane
  (MinDist), which corresponds with the face of the pallet where the merchandise
  is stacked;
Split up the point cloud into cells (voxels), and define the point cloud with points
  into the centroids of the cells (CS = CellSide);
while all the points of the original point cloud have not been processed yet do
    Search the point cloud for the point with the highest distance to the ground
      plane which has not been processed yet (CPH = CurrentPointHeight);
    Add the volume of the prism corresponding to the cell of this point:
        $V+ = CS^2 * (CPH - MinDist)$        (3)
    Mark the cell as processed, as well as all the cells placed underneath (all the
      cells which have been taken into account with the prism)
end
```

The proposed algorithm starts from the idea of splitting up the point cloud into cubes and then adding to the volume sets of cubes, or in other words, add prisms, resulting in a sort of volumetric *integration*. The steps of this algorithm to estimate the volume V, starting from the ground plane and the point cloud obtained in the 3D registration which depicts the merchandise, are described in algorithm 2.

5.2. Method of projecting planes

The previous method is valid for all type of loads. However, one of the most usual ways of merchandise transportation is through boxes. For this reason, the experimental trials have been done with boxes. As long as we accept the hypothesis that all the load is packed in this manner, which is quite plausible, such information can be applied to develop a new volume computation method. Since boxes are formed by planes exclusively, their volume can be calculated through their vertical projection. In particular, the volume of the set corresponds with what there are underneath every plane parallel to the floor, as it can be observed in Fig. 17. Therefore, it is only needed to know all the planes of the point cloud and compute the area of those parallel to the floor. By knowing their height in relation to the pallet, the volume can be easily computed. The steps for computing the volume V of the merchandise represented in a point cloud are explained in algorithm 3.

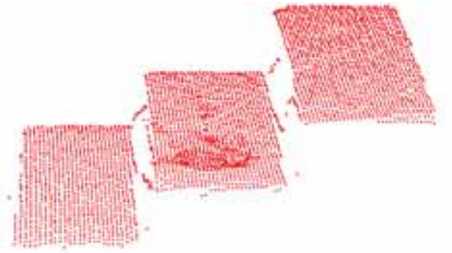


Figure 17: Mesa's point cloud acquired from the top of the arch, depicting a zenithal view. This configuration of the merchandise corresponds with the case 14 of Fig. 19 and consists of three columns of boxes with different heights.

One difficulty of this algorithm is the area computation. The reason is that sometimes the planes have irregular shapes and concave zones. Formulas for area computation of irregular polygons, assuming ordered vertices, exists. However, such information is not known a priori.

The adopted solution consists in the transformation of the 3D point cloud into a 2D image (left picture in Fig. 18, which represents the situation in which the two upper planes of the load of Fig. 17 have been projected into

Algorithm 3: Volume estimation with method of *projecting planes*

```
input  : 3D registration point cloud, a Mesa SR 4000 point cloud (zenithal view)
         and the equation of the ground plane.
output: Volume of the merchandise.
Search the 3D registration for the point with the minimum distance to the floor
  plane ($MinDist$), which corresponds with the face of the pallet where the load is
  stacked;
Search the point cloud acquired by Mesa SR 4000, which represents a zenithal
  view, for all the planes;
Order them according to their distances to the origin (the camera itself);
while all the found planes have not been processed yet do
    Extract the closest plane to the origin (reference plane) which has not been
      processed yet from the point cloud ($RPH = ReferencePlaneHeight$);
    Compute its area ($A = Area$);
    if there is any plane (current plane) beneath the reference one then
        Compute the volume until this lower plane ($CPH = CurrentPlaneHeight$):
                    $V+ = A * (RPH - CPH);$                              (4)
        Project the points of the reference plane onto the current plane;
    else
        In case there are not any planes below, compute the volume until the base:
                    $V+ = A * (RPH - MinDist);$                          (5)
    end
    Mark the reference plane as processed;
end
```

the height of the lowest plane). After that, the image is processed by means of computer vision techniques. First, algorithms such as *eroding* and *dilating* are applied. Note that the second one adds new points around those existing already. As a consequence, the shift within the image is minimum. However, the outlines of the image are displaced (i.e. an image dilation). The effect of the other algorithm is just the contrary. Therefore, unless there are holes or not connected zones in the image which would have been reduced or even disappeared, the final result is practically the original image. The utilization of these functions improves the robustness and precision of the contour detection carried out next. Once we dispose of a connected image, a Canny filter is applied [36] to get the contours of the image and afterwards order them (right picture in figure 18). Finally, the pixels within each contour are counted by means of a function based on the *Green's theorem* [37] to compute the area:

$$A = \oint_{\partial D} P(x,y)\mathrm{d}x + Q(x,y)\mathrm{d}y = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \mathrm{d}x \; \mathrm{d}y \; , \tag{6}$$

where D is a region with boundary ∂D, P and Q functions of (x, y) defined on an open region containing D, where they have continuous partial derivatives. The left side is a line integral and the right side is a surface integral. Manipulating this equation, a connection between the area of a region and the line integral around its boundary is obtained. For a plane curve specified parametrically as $(x(t), y(t))$ for $t \in [t_0, t_1]$, (6) becomes:

$$A = \frac{1}{2} \int_{t_0}^{t_1} (x\ y' - y\ x')\ \mathrm{d}t\ . \tag{7}$$

The area computed in this way is measured in pixels, so it must be transformed to m^2 applying the scale factor between the 2D image and the 3D point cloud.

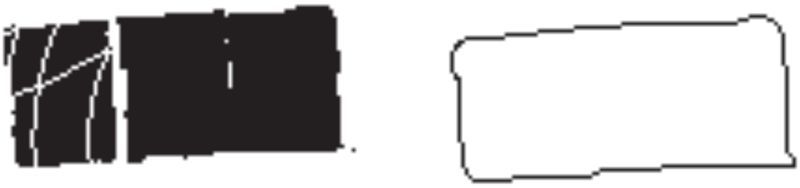
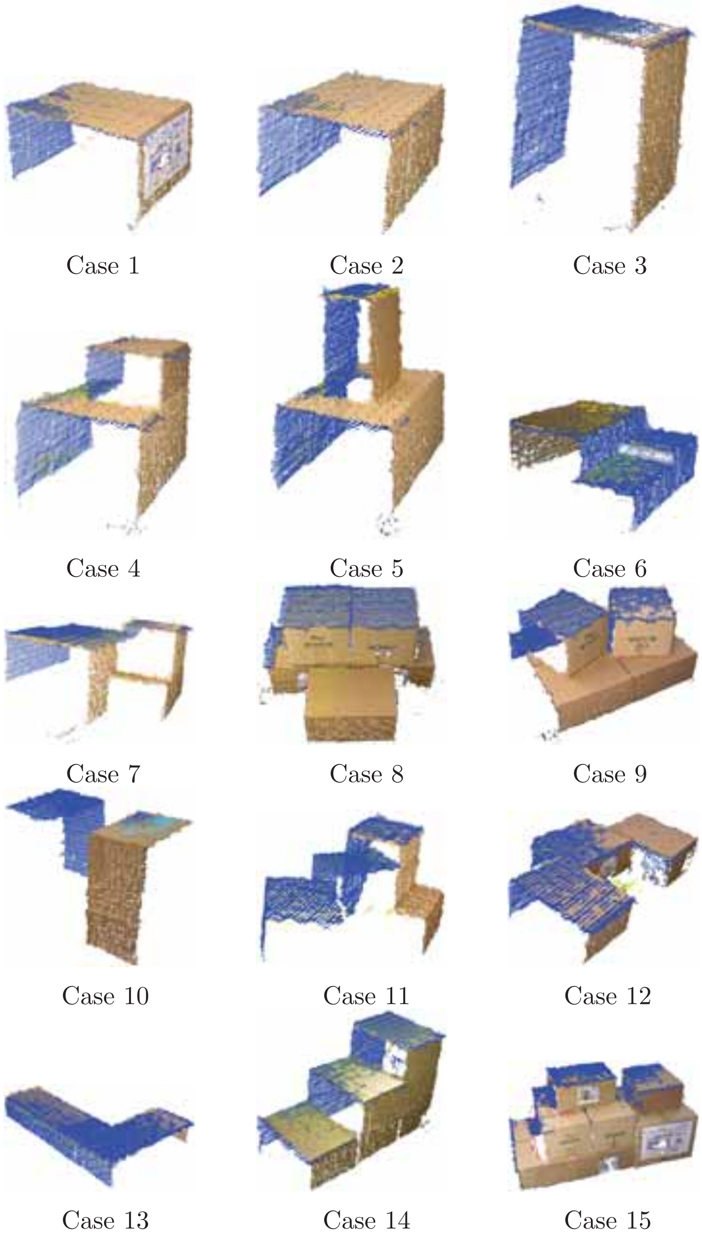


Figure 18: The left picture corresponds to a top view of the point cloud shown in Fig. 17 after projecting the two upper planes onto the plane below (the three planes of this merchandise have been projected, so all points have the same height). On the right, it is displayed the same picture after being processed applying a *dilating* and an *eroding* filter, then a *Gaussian* filter and afterwards the *canny* edge detector. The area of the plane is computed through this image.

6. Experimental results

Once every phase which composes the designed method has been described, we proceed now to evaluate the volume estimation accuracy. A diagram summarizing the different components of the system from the software point of view is depicted in Fig. 20. It also shows the links between the libraries used and the cameras and the computer. Key components of this architecture are the drivers for the data acquisition and the Point Cloud Library used to implement our method relying on different algorithms for point cloud processing.



Case 1 Case 2 Case 3

Case 4 Case 5 Case 6

Case 7 Case 8 Case 9

Case 10 Case 11 Case 12

Case 13 Case 14 Case 15

Figure 19: 3D registration point clouds of the 15 different cases employed to evaluate the system. It is shown the point cloud of Asus Xtion Pro in blue, the corresponding point cloud of the Asus Xtion Pro Live is coloured and the point cloud acquired by Mesa SR 4000 is displayed in an intensity scale (where red is the lowest and blue the biggest intensity level).

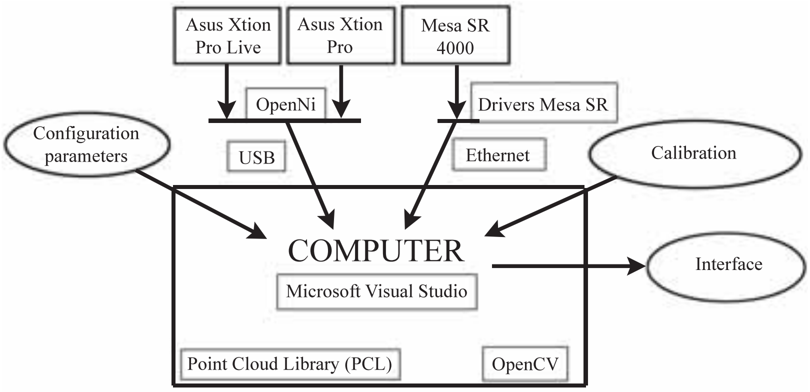


Figure 20: Diagram of the different components involved in the system. Three cameras are connected and controlled with their corresponding drivers. PCL is mainly used for the data processing together with OpenCV library. The output of the system is finally displayed on a TV screen through an interface.

In order to evaluate the volume estimation accuracy, fifteen different trial situations (Fig. 19) have been defined in which the configuration of the merchandise varies from simple situations just with a few boxes to others more complicated with plenty of boxes placed in distinct positions. For every situation, it has been measured the accuracy and computation time applying the two methods previously exposed. Although the trials have been performed with static merchandise, the system can also work when the merchandise is in motion as it can be checked in the **video** attached (the accuracy of the system can be lower in a dynamic situation than in a static one because data quality might decrease). In the video it is displayed how an operator passes through the arch with a load of merchandise and the system returns its volume and 3D registration on a screen. The experiments have been carry out with a 64 bit computer with six Intel(R) Xeon(R) CPU X5650 (2,67 GHz) cores and 24 GB of RAM memory.

A result summary for both volume estimation methods previously explained is shown in the Table 2. On the other hand, in Table 2 we can observe how the error produced with the *voxel* method is quite acceptable. Specifically, the average error is 7.41%, so it is inside the admissible limits for an industrial system. The computation time employed for the volume estimation varies from 1 to 5 seconds, depending on the size of the transported load. In case the merchandise is restricted to boxes exclusively, the method of *projecting plane*s is more efficient, accurate and also with the smallest computation times (around 1 second on average). Furthermore, the resultant average error is only 4.16%.

In order to illustrate the proposed method, a summary of the intermediate steps during a volume estimation process for one of the situations included in the system evaluation (case 9) is shown in Fig. 21. The measured volume for this configuration is 67.5 dm^3, while the estimation of the *voxel* method is 72.5 dm^3 and 69.3 dm^3 with the method of *projecting planes*, which correspond with an error of 7.33% and 2.69% respectively.

Regarding the total time of a scanning, data can be acquired at a frame rate of 30 or 50 frames per second (see Table 1). The scene segmentation and 3D registration phases only require a few seconds and their processing time is quite stable. Thus, the most relevant stage in terms of efficiency is the volume estimation algorithm as shown in Table 2. On average, the total time of scanning from the initial acquisition until the final volume estimation shows up in the user's screen is about 8 seconds. Therefore, the system can be considered to work in real time since the merchandise does not need to be stopped to wait for the result of the estimation. Notice that the code has not been optimized yet and the full hardware capacity is not exploited. For instance, apart from the volume estimation, additional data is shown in the screen during the experiments in live such as the merchandise reconstruction, different steps of the algorithm or information about the characteristics of the merchandise.

Table 2: Summary of the results obtained for volume estimation with each method for the Fifteen Cases. Time measurements refer to the computation time of the volume estimation algorithms. A positive error corresponds with an overestimation of the volume, while a negative error is an underestimation.

Case	Voxel		Projection of planes	
	Error (%)	Time (s)	Error (%)	Time (s)
1	7.37	1.14	-2.47	0.43
2	4.30	1.19	-2.86	0.39
3	10.11	0.83	0.52	0.42
4	9.83	1.77	4.16	0.94
5	7.92	1.95	2.19	1.12
6	-0.12	2.21	-4.60	0.76
7	10.13	2.06	3.65	0.80
8	9.40	1.92	7.97	0.92
9	7.33	2.43	2.69	0.95
10	7.65	0.66	4.86	0.27
11	7.00	1.98	-0.76	1.11
12	9.97	1.63	-7.07	0.57
13	-1.41	2.44	-13.24	0.83
14	11.73	2.05	5.08	1.24
15	6.81	5.56	-0.23	1.81
Mean error	**7.41%**		**4.16%**	
Median error	**7.65%**		**3.65%**	

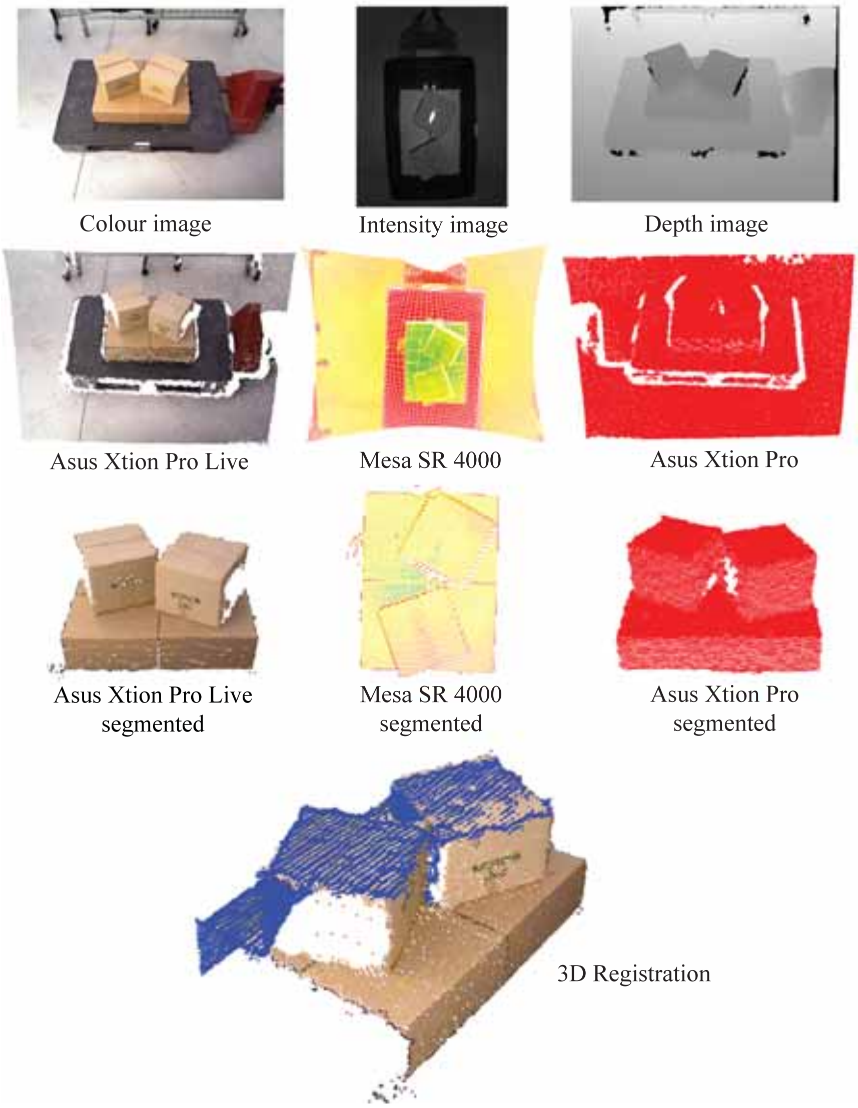


Figure 21: Summary of the intermediate steps for the volume estimation process in the case 9. The first row corresponds with the raw image acquired by the devices. The second row is the same image after being transformed into a point cloud. The third row represents the point cloud after being segmented, that is to say, only the merchandise remains. Finally, the last point cloud is the result of doing a 3D registration with the segmented point clouds. Starting from this point cloud, the final step is the volume computation of the merchandise through the application of the designed algorithms.

7. Conclusion

Given the obtained results, it can be stated that a robust and efficient system has been designed, which is able to estimate the volume of the merchandise stacked on a pallet. Besides, the challenge that entails working with sensors based on distinct technologies, such as structured light or time of flight, has been overcome. The developed system has been experimentally evaluated with the volume estimation in fifteen disparate situations which cover a wide variety of load configurations. The results show that the developed system is accurate, with a margin of error smaller than 10% on average and can be performed in real time. Although this accuracy error could be reduced by improving some particular aspects of the implemented system, the margin of improvement is quite limited because of the measurement errors inherent to working conditions and the own devices as it has been observed in the accuracy and precision study.

Concerning to the cost of the measurement system, it depends on the number and type of sensors used. For instance, one Asus can costs approximately about 150 € and one Mesa about 4.000 €. Note that this field of research is growing very fast and the costs of these sensors may be reduced in medium-term period of time. Additionally, a computer is needed. Although it seems preferable to choose the Asus, it must be taken into account that it is not an industrial device. Therefore, it carries a potential lack of robustness and, for this reason, it would not be recommendable to use this sensor in some specific environments with hard conditions such as vibrations or environmental dust.

Given the good results obtained in the experimental evaluation, the Technological Institute of Aragón[1] has installed the proposed system in the ICT Logistics Demonstration Centre. There, the application is exhibited to companies which visit these installations in order to show them the possibilities that range sensors offer in conjunction with the open-source library PCL.

Future work for enhancing the features of the proposed system could be the combination with other technologies for the purpose of getting more detailed information of the merchandise or even to improve volume estimation. A suitable technology is Radio-frequency identification (RFID), which is being increasingly applied in environments where our system can be implemented.

[1] http://www.itainnova.es

Acknowledgment

The authors would like to thank Guillermo López (ITA INNOVA) his assistance to the execution of this work. The work has been partially supported by the Ministerio de Economía y Competitividad of Spain (DPI2014-61792-EXP and DPI2015-65962-R) and by the Technological Institute of Aragón (grant PFCLOGISTICA1-13-458).

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, vol. 92, pp. 119–136, 2009.

[2] M. Omid, M. Khojastehnazhand and A. Tabatabaeefa. *Estimating volume and mass of citrus fruits by image processing technique*. Journal of Food Engineering, vol. 100, pp. 315–321, 2010.

[3] B. Upcroft, R. C. Shekhar, A. J. Bewley, S. Leonard and P. J. A. Lever. *Measurement of bulk density of the payload in a dragline bucket*. US20120136542 A1, 2012.

[4] A. J. Bewley, R. C. Shekhar, S. Leonard, B. Upcroft and P. J. A. Lever. *Real-time volume estimation of a dragline payload*. IEEE International Conference on Robotics and Automation (ICRA), pp. 1571–1576, 2011.

[5] E. Duff. *Automated Volume Estimation of Haul-Truck Loads*. Proceedings of the Australian Conference on Robotics and Automation, pp. 179–184, 2000.

[6] Y. Takahashi, T. Hakucho, C. Miyake, B. Jiang, D. Morimoto, H. Numasaki, Y. Tomita, K. Nakanishi and M. Higashiyama. *Diagnosis of Regional Node Metastases in Lung Cancer with Computer-Aided 3D Measurement ofthe Volume and CT-Attenuation Values of Lymph Nodes*. Academic Radiology, vol. 20, pp. 740–745, 2013.

[7] A. Chromy.*High-Accuracy Volumetric Measurements of Soft Tissues using Robotic 3D Scanner*. 13th IFAC and IEEE Conference on Programmable Devices and Embedded SystemsPDES, vol. 48 pp. 318 - 323, 2015.

[8] B. Dellen and I. A. Rojas. *Volume measurement with a consumer depth camera based on structured infrared light.* 16th Catalan Conference on Artificial Intelligence, 2013.

[9] D. J. Lee, J. Eifert, P. Zhan and B. Westhover. *Fast surface approximation for volume and surface area measurements using distance transform.* Optical Engineering, pp. 2947-2955, 2003.

[10] S. Seitz, B. Curless, J. Diebel, D. Scharstein and R. Szeliski. *A comparison and evaluation of multi-view stereo reconstruction algorithms.* Proc. CVPR, pages 519-528, 2006.

[11] B. Tippetts, D. J. Lee, K. Lillywhite and J. Archibald. *Review of stereo vision algorithms and their suitability for resource-limited systems.* Journal of Real-Time Image Processing, pp. 1–21, 2013.

[12] L. Cruz, D- Lucio and L. Velho. *Kinect and RGBD Images: Challenges and Applications.* Conference on Graphics, Patterns and Images Tutorials (25th SIBGRAPI-T), pp. 36–49, 2012.

[13] A. Kolb, E. Barth, R. Koch and R. Larsen. *Time-of-Flight Sensors in Computer Graphics.* Proceedings of Eurographics, pp. 119–134, 2009.

[14] M. Hansard, S. Lee, O. Choi and R. Horaud. *Time of Flight Cameras: Principles, Methods, and Applications.* Springer Briefs in Computer Science, November, 2012.

[15] S. Song and J. Xiao. *Tracking Revisited Using RGBD Camera: Unified Benchmark and Baselines.* Proceedings of the 2013 IEEE International Conference on Computer Vision, pp. 233–240, 2013.

[16] H. Shen, B. He, J. Zhang and S. Chen. *Obtaining four-dimensional vibration information for vibrating surfaces with a Kinect sensor.* Measurement, vol. 65, pp. 149–165, 2015.

[17] P. Meißner, S. R. Schmidt-Rohr, M. Lösch, R. Jäkel and R. Dillmann. *Localization of furniture parts by integrating range and intensity data robust against depths with low signal-to-noise ratio.* Robotics and Autonomous Systems, vol. 62, pp. 25–37, 2014.

[18] Z. Xue, S. W. Ruehl, A. Hermann, T. Kerscher and R. Dillmann. *Autonomous grasp and manipulation planning using a ToF camera.* Robotics and Autonomous Systems, vol. 60, pp. 387–395, 2012.

[19] R. B. Rusu and S. Cousins. *3D is here: Point Cloud Library (PCL).* IEEE International Conference on Robotics and Automation (ICRA), China, pp. 1–4, 2011.

[20] J. Fischer, G. Arbeiter, and A. Verl. *Combination of time-of-flight depth and stereo using semiglobal optimization.* IEEE International Conference on Robotics and Automation (ICRA), pages 3548–3553, 2011.

[21] Y. Kim, D. Chan, C. Theobalt, and S. Thrun. *Design and calibration of a multi-view TOF sensor fusion system.* In Proc. CVPR Workshop on time-of-flight Camera based Computer Vision, 2008.

[22] S. Rogge and C. Hentschel. *A multi-depth camera capture system for point cloud library.* International Conference on Consumer Electronics - Berlin (ICCE-Berlin), 2014.

[23] T. Stoyanov, R. Mojtahedzadeh, H. Andreasson and A. J. Lilienthal. *Comparative Evaluation of Range Sensor Accuracy for Indoor Mobile Robotics and Automated Logistics Applications.* Robotics and Autonomous Systems, vol. 61,pp. 1094–1105, 2013.

[24] R. Martin-Martin, M. Lorbach and O. Brock. *Deterioration of depth measurements due to interference of multiple RGB-D sensors.* Intelligent Robots and Systems (IROS 2014), pp. 4205-4212, 2014.

[25] H. Gonzalez-Jorge, B. Riveiro, E. Vazquez-Fernandez, J. Martínez-Sánchez and P. Arias. *Metrological evaluation of Microsoft Kinect and Asus Xtion sensors.* Measurement, vol. 46, pp. 1800–1806, 2013.

[26] *The International Vocabulary of Metrology Basic and General Concepts and Associated Terms (VIM)*, 3rd edition. JCGM 200:2012.
http://www.bipm.org/en/publications/guides/vim.html.

[27] A. Nguyen and B. Le. *3D point cloud segmentation: A survey.* 6th Robotics, Automation and Mechatronics (RAM), pp. 225–230, 2013.

[28] U. Weiss and P. Biber. *Plant detection and mapping for agricultural robots using a 3D LIDAR sensor*. Robotics and Autonomous Systems, vol. 59, pp. 265–273, 2011.

[29] M. A. Fischler and R. C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Commun. ACM, pp. 381–395, 1981.

[30] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, 1990.

[31] G. K. L. Tam, Z. -Q. Cheng, Y. -K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X. -F. Sun and P. L. Rosin. *Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid.*, IEEE Transactions on Visualization and Computer Graphics, pp. 1199-1217, 2013.

[32] K. Pathak, A. Birk, N. Vaškevičius and J. Poppinga. *Fast Registration Based on Noisy Planes With Unknown Correspondences for 3-D Mapping*. IEEE Transactions on Robotics, vol. 26, pp. 424–441, 2010.

[33] Z. Langming, Z. Xiaohu and G. Banglei. *A flexible method for multiview point clouds alignment of small-size object*. Measurement, vol. 58, pp. 115–129, 2014.

[34] 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., pp. 698–700, 1987.

[35] P. J. Besl and N. D. McKay. *A Method for Registration of 3-D Shapes*. IEEE Trans. Pattern Anal. Mach. Intell, pp. 239–256, 1992.

[36] J. Canny. *A Computational Approach to Edge Detection*. IEEE Trans. Pattern Anal. Mach. Intell., pp. 679–698, 1986.

[37] W. Kaplan. *Green's Theorem*. §5.5 in Advanced Calculus, 4th ed. Reading, MA: Addison-Wesley, pp. 286–291, 1991.