ISSN 2176-5480

22nd International Congress of Mechanical Engineering (COBEM 2013)
November 3-7, 2013, Ribeirão Preto, SP, Brazil
Copyright © 2013 by ABCM

DETECTION OF MOVING OBJECTS FOR ROBOTIC MANIPULATION APPLICATIONS IN NON-STRUCTURED ENVIRONMENTS

Guilherme Fernandes
Engineering School of São Carlos, University of São Paulo, São Paulo, Brazil
gfernandes@usp.br

Leonardo Marquez Pedro
Giovanna Pisicchio Zanoni
Center for Exact Sciences and Technology, Federal University of São Carlos, São Paulo, Brazil
lmpedro@ufscar.br,giovannapzanoni@hotmail.com

Glauco Augusto de Paula Caurin
Engineering School of São Carlos, University of São Paulo, São Paulo, Brazil
gcaurin@sc.usp.br

Abstract. *Robotic manipulation tasks, particularly in unstructured environments, require detailed information about the manipulator surroundings. Location and form of objects to be manipulated are especially important information. Due to the emergence of RGB-D cameras, e.g. the Microsoft Kinect, the research community w as able to develop important results related to 3D environment mapping, scene segmentation and object detection. Calibration between two cameras, depth camera and RGB camera provided in the RGB-D devices has been theme of an extensive number of papers. Nevertheless not only the intrinsic and extrinsic cameras parameters are necessary, the transformation between the reference frames of the camera and the reference frame of the manipulator is mandatory information for any arbitrary application. Frequently this measurement procedure is difficult to perform and implies in parameter values deviations. This paper presents an experimental method to obtain the transformation parameters for the two reference frames. The robot modeling is elegantly developed using the product of exponentials approach and the twist between robot and camera reference frames which is obtained, and represents the best solution in a least squares sense. Experimental application using a SCARA 4 D.O.F robot and a Kinect RGB-D camera is conducted to ascertain the method effectiveness. This approach represents an improvement and a flexibilization for robot programming tasks.*

Keywords: *3D Vision, Robotic Manipulation, Camera Calibration, Robot Localization*

1. INTRODUCTION

In the field of robotic manipulation, the use of a vision system is essential to identify the geometric constraints imposed by the robot structure and the objects contained in the environment. The importance of a vision system is better perceived if the human vision is taken as a reference. The human visual-motor coordination process begins with the image acquisition made by the eyes followed by several cognitive processes performed by the central nervous system (CNS). Identification of lines and borders (Zeki (1993)), object identification and classification and optimized grasping strategies (MacKenzie and Iberall (1994)) are recognized as processes that take place inside the CNS. A classical survey by Cutkosky (1989) suggests that humans use predominant hand postures to pick up objects. The human vision system is used to define the position of objects, to classify them and it helps further to estimate other important properties such as surface texture, density, mass, volume and moment of inertia. Vision also allows the construction of environment model, e.g., a few seconds observation environment are sufficient for a virtual map is constructed by the CNS, containing information such as presence and classification of objects, people, references, obstacles, etc.

A continuous update of the environment conditions with the detection of dynamic changes is provided by the human vision system, in this way the approach of moving objects can be perceived. Visual stimuli are first detected peripherally and they are processed faster than objects classification or environment mapping.When an object is approaching fast, the information follows a different path in the CNS allowing it to respond fast, allowing humans to avoid or mitigate the effects of a collision. This reactive response occurs usually inside 200ms to 300ms without theo bject being classified or identified (Zeki (1993) and MacKenzie and Iberall (1994)).These motor-visual coordination characteristics are relevant for both the design and the implementation of robotic manipulation systems.

State of the art robot manipulation systems relies on the use of 2D vision sensors (CCD cameras) and 3D vision sensors (depth range sensors). The choice of individual solutions or the combination of different sensor with data fusion is defined by the specific application requirements.The detection of moving objects and collision avoidance, for example, require fast response time. In this case, the large amount of data provided by 3D sensors in a single scene acquisition and the corresponding high computational cost represents a disadvantage.

ISSN 2176-5480

In the implementation ofmanipulation systems, 2D based vision technologies are preferred over 3D systems due to their lower processing costs, and better image processing libraries available libraries. There are algorithms implemented for various programming languages, such as OpenCV.

Therefore, 2D systems are adopted when robustness and speed are required by the implemented applications, namely motion detection and segmentation.Conversely, 3D sensors provide better data quality for the same acquisition interval. They may be more useful for higher-level control functionalities such as volumetric information, the identification of object handles or especial and useful environment features that are required for mobile robot navigation purposes.

This work focuses on robotic manipulation applications that take place inside unstructured environments. In this context, the robot will often contact objects, and humans. The computer vision system proposed here needs to detect moving objects for different reasons: avoiding and treating collisions, giving support to contact handling, object classification, and more importantly providing input to grasping and manipulation procedures.

When a vision system is used it is necessary to establish a mathematical relation between the sensor measurement coordinates and the robot sensor coordinates. In this way it is possible to relate both information. A method is presented in this paper deriving and identifying the parameter of the homogeneous transformation matrix relating sensor data to robot reference frame.

In the sequence, a method is implemented for the detection of dynamic objects. The approach is based on RGB images using conventional imaging processing algorithms that present rapid processing characteristics. As the adopted 3D sensor present a registry relating range and RGB images, detection performed on the RGB image can be used as a mask for the selection of all 3D data belonging to the moving object. Further based on this approach it is possible to estimate position, volume and even classify the object or to perform some required treatment on the corresponding 3D data.

As the robot structure also moves, being able to filter points in the image belong to the robot, make the object detection algorithm more efficient. A simplified segmentation method based on a polygonal model of the robot is provided in this paper.

2. ROBOT AND VISION SENSOR CALIBRATION

2.1 MANIPULATION ROBOT

The experimental procedure was conducted by a 4 degrees of freedom SCARA robot. The original IBM7545 robot have just been retrofitted by replacing its sensors (encoders and limits switches) and also replacing its servo drivers. The new servo drivers adopted were selected intending to provide a motor high-level position, velocity and current control. The EPOS2 70/10 servo drivers from Switzerland MAXON were selected. The communication is conducted by a CAN network over the CANOpen protocol according the CiA CANopen specifications (DS-301 Communication Profile for Industrial Systems, Version 4.02 and DSP-402 Device Profile for Drives and Motion Control, Version 2.0). The CANOpen implementation granted the real-time architecture capacity by its reliability and determinism.

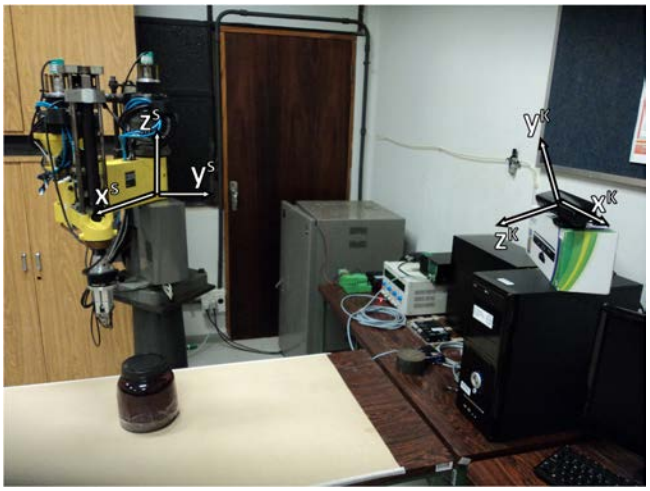


Figure 1. SCARA robot and Kinect sensor.

2.2 KINECT SENSOR

Cost reduction is always a requirement to be meet in the selection of an equipment, mainly in laboratorial environments. Therefore developers from different areas have been adopting the Microsoft Kinect, originally developed with for entertainment purposes, as range imaging sensor thanks to its low cost to replace the expensive laser scanner systems in

ISSN 2176-5480

applications that do not require long distances and high positioning accuracy. The use of the sensor for these applications began after it was launched in late 2010, which, prompted Microsoft to publish a Software Development Kit in February 2012 Pedro and de Paula Caurin (2012).

The Kinect sensor has a digital camera, with 640x480 pixels of resolution, and a range imaging sensor, with 320x240 points of acquisition. Both have 30 frames per second as acquisiton rate. It is noteworthy that, with some restrictions, there is a correspondence between the RGB image and range image, i.e. which composes a RGB-D image.

The Kinect depth sensor operation is based on a triangulation process. A transmitter sends an infrared laser beam that is diffracted by a projection lens into multiple beams. The pattern of these beams projected on the scene is then captured by a camera that is only sensitive to wavelengths in the infrared. The pattern of captured beams are then compared with the standard beam obtained from the calibration of the sensor. By comparison of the pattern obtained in each acquisition with the pattern from a the sensor calibration, it is possible to determine the x, y and z coordinates of each pixel. This produce an array of three-dimensional points. A more detailed description is presented by Khoshelham (2010).

The results of Kinect repeatability and precision experiments presented by Khoshelham (2010) and Pedro and de Paula Caurin (2012) suggest that the sensor meets requirements for data acquisition at distances between 1m and 3m, with an accuracy ranging from 5 to 10mm respectively.

2.3 ROBOT AND SENSOR CALIBRATION

The SCARA robot and the Kinect sensor have different reference systems. The robot has a system fixed on S its base, as shown in Figure 1, in which the position of a point P representing its tool center point can be represented by coordinates x, y and z as $P^S = [x^s\ y^s\ z^s]^T$, and the sensor has a reference system K according to Figure 1. The calibration between the SCARA and Kinect consists on finding the matrix homogeneous transformation H_K^S such that:

$$P^S = H_K^S P^K \tag{1}$$

The matrix H_K^S between Kinect and SCARA reference system is necessary to describe the 3D position of each sensor acquisition points in the robot reference system. Once it is not intended to fix statically the sensor in relation to the robot, the matrix determination must be performed whenever the sensor is repositioned, or before specific applications to ensure that the data obtained by the sensor are properly transformed into the robot workspace. Thus, the sensor can be repositioned depending on the requirement of each specific application serving diverse applications beyond this study.

A method for determining the matrix consists on the Kinect measurement of identification points, composed by a LED, fixed to the robot. The measurement method is that presented by Pedro and de Paula Caurin (2012). For each position i of different robot joints positions, both robot position, represented by $P_i^S = [x_i^S\ y_i^S\ z_i^S]^T$, and LED position measured by Kinect, represented by $P_i^K[x_i^K\ y_i^K\ z_i^K]^T$ are recorded for further calculation. Once n points are measured, it is possible to construct the following system:

$$[P_1^S\ P_2^S\ ...\ P_n^S] = H_K^S[P_1^K\ P_2^K\ ...\ P_n^K] \tag{2}$$

This system has no exact solution due to Kinect measurement errors and quantization errors when considering that the LED represents a spatial point. The maximum robot positioning error is $0.013mm$, thus, the observed errors are from the sensor and from the identification points measurements procedure Pedro and de Paula Caurin (2012). To solve the system it is addopted the method of least squares to find a solution of linear systems in the form $Ax = B$ by minimizing the deviation $|Ax - B|$.

By means of matrix operations it can be shown that the equation:

$$(P^K)^T(H_K^S)^T = (P^S)^T \tag{3}$$

is equivalent to equation 1, but written in the form $Ax = B$, where $(P^S)^T = B$, $(P^K)^T = A$. This linear system can be solved by Matlab with the command " » mldivide(A,B)".

Measuring identification points with Kinect for different positions of the robot it is possible to obtain relations enough to determine the H_K^S.

For calibration efficiency, ie, with low value for $|Ax - B|$, it is necessary that A contains the smallest error possible. Once Kinect measurement error is 100 times greater than those SCARA shown in positioning, which positioning error of its TCP is $0.5mm$). Thus, the identification points for Calibration must be acquired as close as possible to the sensor. And yet, several acquisitions must be performed to a heading of the robot in order to reduce the effects of random errors.

2.4 Evaluation of the Kinect and robot calibration

Experiments were performed to verify the quality of the matrix H_K^S obtained by the method above according to the distribution of calibration points. For exemplification, figure 2 shows the points used in the calibration procedure.

ISSN 2176-5480

After calibration, the identification point was positioned over the entire robot workspace of the robot in $5mm$ spacing. Each of the different positions was measured with Kinect. To eliminate random measurements errors, a total of 30 measures were performed and its medians were considered as the P^K measured value. Posteriorly, the LED position in the robot workspace was estimated by teh relation $P^S_{est} = H^S_K P^K$. Finally, the estimated position was compared with the real position P^S_{real}.

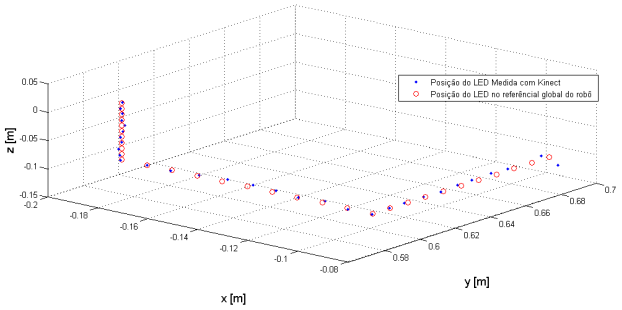
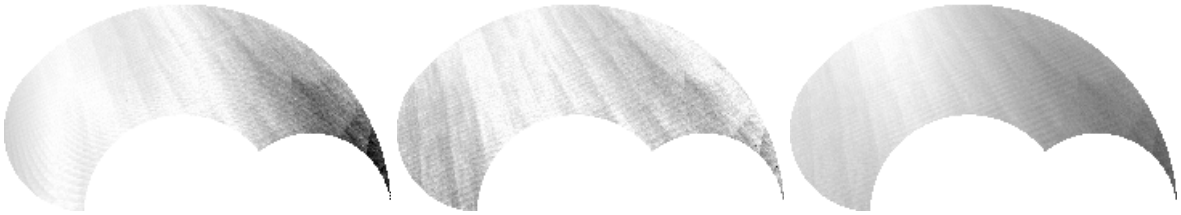


Figure 2. Calibration points.

The error transformation given by the difference between P^S_{real} and P^S_{est} is represented in the following figures. Each one of its pixels represents a checkpoint, and its error is represented in grayscale. White pixels have zero error while black pixels represent errors of $40mm$ or higher. figures 3(a), 3(a) and 3(a) show the transformation error for x, y and z coordinates, respectively.



(a)

Figure 3. Calibration error represented in the robot workspace. Figure 3(a) shows the errors for x coordinate, figure 3(a) shows the errors for y, and figure 3(a) shows the errors for z coordinate

3. Moving objects detection

The detection of moving parts is performed on the RGB image Kinect. The result of this detection is a binary image with values equal to one (1) to the pixels belonging to the identified object and zero (0) for pixels belonging to the background scene. The binary image of moving objects is used as a mask over the range image to extract 3D points of the objects.

Then the vision system provides as output a binary image, and a cloud of points of the detected moving objects. Both data can be used later on diverse functionalities of robotic manipulation system, especially in the context of this work, moving object information is intended to implement collision avoidance methods.

3.1 Background subtraction methods

In the literature, there are several proposed methods and algorithms for image background subtraction. Due to the large amount of methods and this topic is not the focus of this work, it is not the purpose of this section to bring a full review of the methods and its details. A review of the main methods of extracting and background is based on a qualitative comparison results obtained by Piccardi (2004)) in order to support and justify the choice of the method selected for image background subtraction, i.e. moving object detection.

Background subtraction of scenes is a procedure used to refer to the moving regions in a sequence of images obtained by a static camera. The main idea is to detect changes between the current image and a background reference, which should be free of moving objects and must be updated as adaptation to variations in brightness and intrinsic changes of scene as shadows, trees and lakes, rivers or flags surfaces waving in the wind.

Most work in this line focus efforts on the development of methods for updating image reference requiring little computational resources such as processing and memory, and are efficient the modi- tions of the scene such as those mentioned above. After updating the reference image, the background subtraction scene can be simplified, for example, a pixel by pixel comparison logic, or even more complex considering statistical measures of variation of color at each pixel.

Piccardi (2004)) gives a brief description and a qualitative comparison of the following methods: Running Gaussian Average; Temporal Median Filter; Mixture of Gaussians; Kernel Density Estimation; Sequential KD Approximation; Cooccurence of Image Variations; Eigenbackgrounds.

Considering memory and processing requirements, the first two methods are those with lower requirement. Others require more computational resources for implementation, such as Mixture of Gaussians, which has been implemented

ISSN 2176-5480

with parallel processing techniques using the Nvidia graphics card.

For the accuracy and precision of the background subtraction, Piccardi (2004) also provides a qualitative comparison. Second the author, the first two methods show good adaptation to slow changes in lighting, but do not have good performance in the correction of rapid background change. However, when such a change occurs distributed in small areas, the problem can be solved by the use of suitable filters.

The method Temporal Median Filter is chosen as the background subtraction method for implementation of the moving object detection.

4. EXPERIMENTS AND RESULTS

4.1 Moving objects detection results

To identify the moving objects, it was selected the Running Gaussian Average method with which it is possible to detect moving objects in a sequence of images. The images of figure 4 show the result of background subtraction identifying the robot and a second object in the scene. Figure 4(b) shows the scene RGB image, and figure 4(b) shows the resulting image from background subtraction.

(a) (b) (c)

Figure 4. igure 4(b) shows the scene RGB image, figure 4(b) shows the resulting image from background subtraction, and figure 4(c) shows the cloud of points.

The resulting binary matrix of background subtraction is used as a mask on range image acquired by Kinect. Figure 4(c) shows the point cloud of the scene adopted as an example, it is possible to observe the points belonging to the robot, to the object, and noise from both problems related to the method of subtraction, such as shadows and lighting variations, and probles due to Kinect registration errors.

4.2 Robot identification and three-dimensional segmentation

The end result of the detection of moving objects includes the robot itself. It is then necessary to identify among all the detected points, which ones belong to the robot. Thus it is possible to isolate only the objects of interest and implement algorithms collisions avoidance, or algorithms for grasping and manipulation. Besides identifying the points belonging to the robot, it is also important to reduce noise (regardless of its source).

For this identification, it is used a simplified polygonal model representing the robot as shown in figure 5. It is important to emphasize that a more detailed model was discarded since there are considerable errors, since from the accuracy and repeatability of the Kinect, to calibration errors.

5. CONCLUSIONS AND FUTURE WORKS

The system proposed in this work is able to detect moving objects in unstructured environment robotic manipulation tasks. The system is based in artificial vision 3D camera, namely Microsoft Kinect, which is fixed statically towards the robot workspace.

The Kinect sensor captures both RGB and depth images. The calibration of the relation (orientation and translation) between the robot reference frame and the camera reference frame is required in order to use the vision data. This relation results in a homogeneous transformation matrix between the two reference frames. An approach to obtain this matrix was presented and experimental procedures were conducted to evaluate the results accuracy.

The experimental results suggest that the calibration error increases accordingly distance from the sensor. Errors up to 40mm on the sensor boundaries position were measured. The experimental procedure was conducted several times purposeful to avoid repeatability errors. The results summary reports a conjunction of several errors: Kinect precision errors, camera distortions errors, point detection errors and calibration method errors.

In practical terms, calibration errors should be considered together with repeatability and precision errors. Therefore,

ISSN 2176-5480

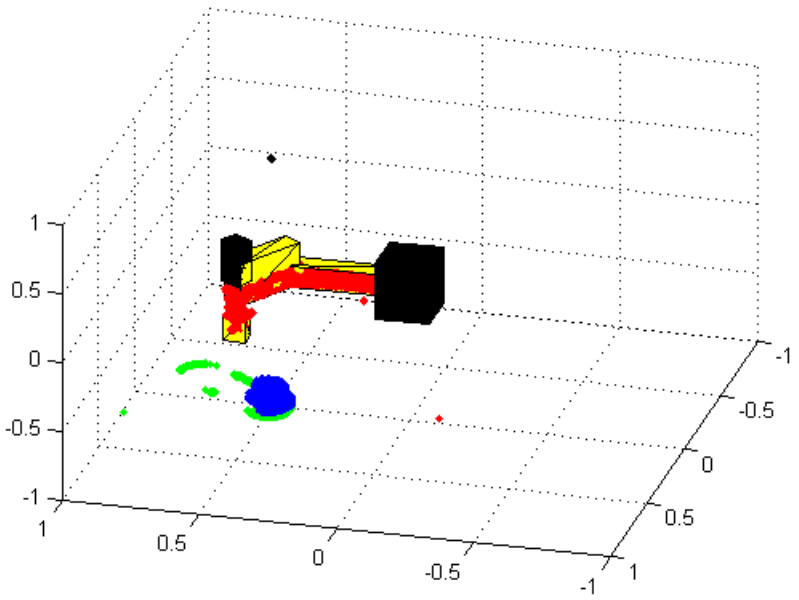


Figure 5. Example of a three-dimensional segmentation. Moving object identified is shown in blue.

in convergence with previous Kinect evaluation results [Pedro] the usage of Kinect as vision information data should be limited to a distance of 1m to 1,5m of the sensor in application with high precision requirements.

Thus the trustable data for grasping and robotic manipulation is the data acquired next to the sensor. However for applications such as collision avoidance e prevention the low accuracy data obtained in higher distances is considered feasible to be used.

Upon the introduction and evaluation of the calibration method, it was detailed a method proposal for image processing for moving objects detection and 3D segmentation. The method objective is gathering a point cloud for the moving object. In order to evaluate the proposal, the image processing method applied to make the background subtraction was the simplest but less effective among all literature methods.

The images arising from the background subtraction algorithm are applied to mask the depth image data. The results are point clouds of moving objects, including the robot itself. The 3D segmentation objective is to split the point cloud in objects belonging and non belonging to the robot structure.

From a simplified polygon model of the robot was possible to detect and qualify the point clouds belonging on their Euclidian distance measurement between the robot model and the point cloud. The results showed is satisfactory even using a simplified robot model. The resultant data is a moving cloud point in the robot reference frame.

6. ACKNOWLEDGEMENTS

The authors thank CNPq and FAPESP for the financial support: CNPq processes numbers 301417/2007-5, 143124/2009-9 and 130222/2012-7; and FAPESP process number 2008/09530-4.

7. REFERENCES

Cutkosky, M.R., 1989. “On grasp choice, grasp models, and the design of hands for manufacturing tasks”. *Robotics and Automation, IEEE Transactions on*, Vol. 5, No. 3, pp. 269–279.

Khoshelham, K., 2010. “Accuracy analysis of kinect depth data”. *GeoInformation Science*, Vol. 38, pp. 1–6.

MacKenzie, C.L. and Iberall, T., 1994. *The grasping hand*, Vol. 104. North Holland.

Pedro, L.M. and de Paula Caurin, G.A., 2012. “Kinect evaluation for human body movement analysis”. pp. 1856–1861.

Piccardi, M., 2004. “Background subtraction techniques: a review”. *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 4, pp. 3099–3104. ISSN 1062-922X.

Zeki, S., 1993. *A Vision of the Brain*. Oxford Univ Press.

8. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.