Abstract—In this paper we present a module for detection and segmentation of human torsos for Human-Robot Interaction. People detection is a very important task in Human-Robot Interaction and particularly at the beginning of this interaction. Visual and 3D data coming from a Kinect sensor is used to achieve the task. The module proposed uses three main process: a) a face detector b) a skin color detector and c) a 3D region growing segmentation technique. Open sources libraries like OpenNI, OpenKinect and Point Cloud Library are used in our implementation to handle the device and sensors data, as well as OpenCV and GLUT to the image processing and display.

I. INTRODUCTION

Collaboration between humans and robots have increased, over the recent years. More and more robots are created to interact with people in different environments, and with different purposes.

Particularly, is in the field of service robots, where this interaction becomes essential. Robots are expected to collaborate with people, and perform activities such as gesture recognition, surveillance tasks, helping elderly, serving as guides in museums or shopping malls and so on. This last point is the object of interest of this work.

One of the main problems, at this stage of robotics research, is still to make robots capable to detect persons without confuse them with any object or characteristic in the environment. There are several ways and sensors to detect persons by an autonomous robot; despite of, it is mainly by image processing techniques that this is done. Face detection [1] is the most common algorithms. Detecting a face increases the probability to detect a person; however there are still false positives in its detection.

In this paper we present the fist part of module to detect and segment persons using a Kinect sensor. Fusing face and skin color detection to remove some of the false positives commonly founded. However, there are still some common errors in this detection. 3D information is required to be certain that what is seen by the cameras is really a person and not, for example, a picture in the wall.

A 3D region growing is applied to 3D data coming from the Kinect range sensor based on a seed obtained by faces detection process. This operation is performed frame by frame. The importance of this module is the integration with a people-tracking module implemented in a robot guide. Integration of this module will offer to users a more robust and natural way of communication.

This paper is organized as follows: In Section 2 we present related works. Section 3 show a brief description of the main libraries used. Section 4 introduces an analysis of several studies of skin detection, and a fusion of the best technique with the face detection technique. In Section 5, we present the segmentation method, using a region growing in 3D. Section 6 shows experimental results of the integration of the previously mentioned methods. Finally, in Section 7 we present the conclusions and future work.

II. RELATED WORK

People detection and tracking are two processes that have been implemented several times. The approach used will depend from the available sensors, just as the objectives of the application, this is, whole human body detection or part-base detection.

Germa et al. [2] developed a module of Human-Robot Interaction in the guide robot Rackham with the objective of keeping it operating in a museum. Using a monocular camera, the robot detects faces of persons and selects the closest for an interaction. Through a touchscreen, the path target is selected, and combining color and shape, skin blobs and frontal face detection with a particle filter, the tracking is performed.

Wu and Nevatia [3] implemented a module to detect and track one or several persons for surveillance purposes. The main advantages of the module is that detected persons can be partially occluded, walking or just standing, just as its operation with a stationary or moving camera, in environments indoors or outdoors with possibly changing illumination. The detection is performed using detectors learned from a set of silhouette oriented features, called features edgelet. Once the several characteristics are detected, the results are combined to create a probability

Universidad Veracruzana
Xalapa, Veracruz, Mexico
beto_ven5@hotmail.com, anmarin@uv.mx, kris_crack@hotmail.com, luisfelipe.marin@ieee.org
model, which can detect persons even partially occluded and reduces detection’s false alarms. The tracking is performed frame by frame.

Muñoz-Salinas et al. [4] used the disparity map of a stereo camera to segment objects from foreground. Once obtained the objects in foreground, a face detector is applied under these regions to create a color model both the face and torso of persons founded. The tracking is applied with a condensation algorithm using the color and depth of the ellipses projected by the colors models of face and torso. Mendez-Polanco et al. [5] implemented a module for detection and tracking humans in an ActivMedia PeopleBot robot with a stereo camera. Recovering the disparity image of the stereo camera, they extract objects of foreground, on which create templates in blobs of similar proportions to the human body. Once labeled the templates to be possible humans, a semi-elliptical contour model is applied under these regions. This model consist of two semi-ellipses representing head and torso’s shape. If the model fits, then assumes that it’s a person. The tracking is performed using the Euclidian distance frame by frame, under and Bayesian fusion approach.

Gonsalves & Teizer [6] detect and track humans using a 3D sensor for surveillance purposes. The tracking algorithm proposed performs the human segmentation from a range image video sequence, using the distance information from the person to sensor. Then, the segmented persons are modeled using a star skeleton structure, and the positions are continuously monitored using a particles filter.

In this paper, we present a module that fuses techniques and 2D/3D information for the detection of possible persons for Human-Robot Interaction, using as sensor a Kinect. The main difference between this proposal with the mentioned above, is the fusion of processing techniques in 2D with processing techniques in 3D, as well as the use of Kinect with the PCL libraries, which are just gaining importance and use. Figure 1 shows the execution sequence of processes in a hierarchical way.

III. SOFTWARE EMPLOYED

As previously mentioned, the present paper uses a Kinect sensor; therefore it is necessary to use libraries for the management of the sensor as well as some others that allow us to manipulate its data. First, we have OpenNI which includes the required drivers for the connection with the device, besides the obtaining of the data structures that it handles; meanwhile Point Cloud Library is a set of libraries that handles the points clouds and 3D geometric processing. Briefly, we will describe the libraries.

A. OpenNI

Open Natural Interaction [7] (OpenNI) is a framework multi-language that defines APIs to write applications using natural interaction. OpenNI provides a set of Application Programming Interfaces (APIs) to be implemented by vision and audio sensors, as well as to be implemented by vision and audio perception middleware.

B. Point Cloud Library

Point Cloud Library [8] (PCL) is a framework to processing point clouds, which includes numerous state-of-the-art algorithms to filtering, feature estimation, surface reconstruction, model fitting and segmentation. PCL handles acquired data through real time devices, as laser scanner, stereo cameras, time-of-flight sensors. PCL is cross-platform and is released under the terms of the BSD license.

IV. 2D PROCESSING

The Kinect sensor provides two types of different structure data: a 2D image and a disparity image. Using the 2D image, characteristics that facilitate the procedure of people detection can be recovered. When looking to identify one person, one of the most commonly used characteristics is face detection. However, this procedure has failed to obtain results completely effective, because always gets false positives, no matter the technique used. In this work, the skin detection technique is fused with the face detection technique, to ensure bigger precision in people detection.

TABLE I

<table>
<thead>
<tr>
<th>Reference</th>
<th>Color Esp.</th>
<th>Accuracy</th>
<th>% Skin</th>
<th>% No Skin</th>
</tr>
</thead>
<tbody>
<tr>
<td>[9]</td>
<td>YCrCb</td>
<td>0.926707031</td>
<td>0.434946325</td>
<td>0.939612385</td>
</tr>
<tr>
<td>[10]</td>
<td>RGB</td>
<td>0.946519594</td>
<td>0.517662652</td>
<td>0.96073312</td>
</tr>
<tr>
<td>[11]</td>
<td>YCrCb</td>
<td>0.949679395</td>
<td>0.440860901</td>
<td>0.963567485</td>
</tr>
<tr>
<td>[12]</td>
<td>HSV</td>
<td>0.964594589</td>
<td>0.190512989</td>
<td>0.986031169</td>
</tr>
</tbody>
</table>

Fig. 1. Hierarchical execution processes
A. Skin Detection

Because the skin detection procedure is not the main objective of this work, was searched for works in the literature about skin detection, which provide the data necessary to replicate his results. Table 1 shows a comparison table with the four main works found about skin detection using the thresholds technique, which reported the used thresholds. In [9] and [11] use YCrCb as space color, getting good results, as we can see in Figure 2 and Figure 3, respectively. Figure 4 shows the results of skin detector [10], however, the results can’t seem better than previous.

B. Face Detection

The face detector most commonly used is the one implemented by Paul Viola & Michaels Jones [13], which use Haar characteristics, based on Haar wavelets. OpenCV implements Viola & Jones algorithm in just one function. Figure 5 shows the result of the face detection process using OpenCV function. As we can see, it can detects more than one person, however, also can confuse and found faces where there’s not.

C. Skin Detection Fused with Face Detection

In order to obtain only true positives, skin detector was applied under the region detected by the face detector module. Using a threshold of skin detected under the region of interest, false positives are filtered. Figure 6 shows the fusion between the two techniques previously mentioned.
V. 3D PROCESSING

Using information from the disparity map from the Kinect sensor is possible to recover the point cloud in 3D. PCL provides function to handle and manipulate 3D data, without the need for sensor calibration. Figure 7 shows a point cloud using Kinect with the PCL libraries.

![Fig. 7. Point Cloud with PCL](image)

The 2D image processing presented above is used to detect human potential in the environment, however, is necessary segment them from the rest of the scene to facilitate the tracking process. This is why we use 3D information obtained with the Kinect sensor, to segment the persons through the distance, implementing a region growing technique, using as seed the geometric center from the bounding box of the face region. Because it is not necessary to obtain the complete silhouette of the person, the region growing ends when it reaches the region of the legs of the person. In this way, we resolve problems of growth in floor area.

To get 3D data from a 2D space, the following conversion is required:

\[ [x, y, z] = \text{cloudPoint}[u + (v \times \text{width})] \]

where \((u, v)\) are the coordinates of pixel or seed in 2D space, width is from point cloud, and \([x, y, z]\) are equivalent to the pixel coordinates in 3D space.

VI. EXPERIMENTAL RESULTS

During the explanation of the methods used, we were showing results of each module independently. This section will show the results of the integration of the modules of face detection with skin detection, using region growing segmentation and tracking frame by frame. Because the detection of persons is done with the face, it is necessary that the person is in front of the sensor.

Figure 8 shows the result of detection and segmentation using a distance threshold for the region growing of 15 cm. As we can see, the region growing does not return the complete silhouette of the person, and ends when it reaches the knee region, approximately. Also we can see that precision is not accurate, because it has noise in the region corresponding to the person adding part of the background. However, the results are acceptable and suitable for the torsos detection process.

Figure 9 shows the tracking process, frame by frame, with more than one person in the scene. As we can see, the distance between people with the sensor does not affect the performance of the module and the region growing algorithm. Because the face detection technique is a very slow process, the obtaining of images and point clouds is affected, running at 2 FPS or so.

VII. CONCLUSION AND FUTURE WORK

In this paper, a part of a module for the detection and segmentation of human torsos has been presented. This module is intended for future use in Human-Robot Interaction. One of the main contributions of this work is the use of data from multiple sources, so that it is able to obtain more accurate results than if we just used one information’s source. The use of libraries OpenNI facilitated the control and data acquisition device, which in this case was the Kinect, while the PCL library enabled faster delivery of 3D data. Another advantage is the easy adaptation to other 3D-type sensor, as for example the SwissRanger sensor, because the data types used are similar in both sensors.

In future work, improvements to the module are:

1. Implementation and integration of a particle filter to add tracking of the torso, and don’t do the detection frame by frame
2. Code optimization to increase frames per second
3. Reduce use of face detection technique for faster process
4. Add other features for detection of the person, so it is not necessary initialization
5. Create a torso model or whole human body model
6. Test with other sensors, like SwissRanger

This module will be integrated with a guide module already implemented in UVerto, which is a robot guide with capabilities like mapping and safe navigation. UVerto uses a face detector to recognize people in the environment, but like we said before, this module may fail and detect persons where there are not. The objectives of this integration is
making UVerto offers their guide services to real people, and also make a robust detection and tracking.

Fig. 8. Detection and Segmentation of Human Torsos

Fig. 9. Detection, Segmentation and Tracking of Torsos Frame by Frame

REFERENCES


