Using Color Histograms and Range Data to Track Trajectories of Moving People from a Mobile Robot Platform

Juan Alberto Rivera-Bautista, Antonio Marin-Hernandez and Luis F. Marin-Urias

Abstract—People detection and tracking is still an active research field for autonomous mobile robots. Approaches that tackle the problem on mobile robots differ from those proposed for surveillance and other similar tasks. Mainly due to the motion of sensors and hence the complexity of scene interpretation, approaches proposed for fixed cameras are not appropriate.

In this paper it is presented a methodology for people detection and tracking from an autonomous mobile robot. People detection and tracking initialization is made by image processing techniques, while for tracking data provided a range data camera (Kinect like sensor) and color information from the former sensor are used to track people over 3D space. Accuracy of tracker is achieved by an adaptive color histograms method based on color from RGB camera and projected over the plain view.

I. INTRODUCTION

Tracking people from a sensor mounted on a moving platform is still a challenging task. There are many available techniques used to track people from fixed cameras (e.g. security cameras). However, due to the fact that on this situation background can be considered as an unchanged characteristic, it can be removed easily [1][2]. This is not the case when using mobile platforms.

Human detection and tracking is one of the most important tasks for autonomous robots intended to share the environment with people. For example, knowing the position and speed of a moving person a robot can modify their motion trajectory to avoid collision or it can compute the human trajectory to identify activities [3].

Nowadays, human robot interaction (HRI) has become an essential part of service robots from museums or office buildings guides to home assistants. In order to do efficiently many of the requested tasks on mentioned environments, a robot must be capable not only detect people on the surroundings, it needs to know the distance to them, as well as theirs speeds and directions. Based on that knowledge, a robot can predict or analyze trajectories followed by people and then propose helping or assistance [4].

As an essential and fundamental part of a service robot, primarily passing by the human audience acceptance, a good and robust module for HRI must be designed. This module begins with an efficient people detection and tracking.

In this paper, the implementation of a real time people detection and tracking module is presented. The proposed module has been designed to work over a mobile robot platform on indoor environments.

In order to obtain robustly results, this module is based on multisensory information. On one side, the processing of the RGB images coming from a monocular camera and on the second the processing of dense depth information proportioned by a range sensor.

The first part of the module provides the existence of humans in the environment and the second offers the position and motion of detected humans in the robot and global reference frame.

The proposed module has been tested with a Kinect like sensor, however it is not limited to this kind of sensors, it can be also implemented over a passive stereo-camera or an active calibrated system formed by a monocular and a time of flight (TOF) cameras.

This paper is organized as follows. We start with a review of the related work on section II.

On section III the development, which lies on image processing, is described followed on section IV by the depth information processing. On section V we present the results and finally on section VI we give our conclusions and future work.

II. RELATED WORKS

People detection and tracking people has been attacked in several ways and with different kind of data and sensors. As has been described by many authors, the use of only one characteristic in complex environments is not enough to do a robust detection and tracking due to inaccuracies and deficiencies on both processes.

The condition that autonomous robots must realize onboard these kinds of task adds complexity.

In order to deal with the mentioned problem, over the last years, many works using multisensory information or multi-method processing have been presented. For example, in [5] is presented a module on a mobile robot to track people, using a laser range finder to recognize leg patterns, and a face detector to fuse the leg detector information, and validate person detection. In [6] is presented a multiple people tracker using a 2D laser-range data and a sample-based joint probabilistic data association filters. The problem of these works is the large number of false positives returned by the legs detection technique.

Beymer and Konoligein [7], present a people tracking
method using a stereo camera and an occupancy grid map that is constructed incrementally by the robot with known SLAM techniques. People are modeled over floor plane as cylinders and the tracking process is doing by the use of a mixture of Gaussians over the re-projection of 3D data over the floor plane. On this work, the camera focal axe is assumed to be parallel to the floor plane, so re-projection is done simply eliminating Z (vertical) component. Over more realistic experiments is not always possible to consider the camera focal axe parallel to the floor plane, mainly due to the proper motion of the robot (e.g. legged robots) or to small differences between floor tiles that makes undesirable motions on wheeled robots.

With the recently appearance of cheap range data sensors as Microsoft’s Kinect some works has been proposed dealing with people detection and tracking using this kind of sensors. However they are still restricted and the use or implementation over mobile robot platforms is limited.

For example in [8], is presented a method for human pose estimation using this sensor without using temporal estimation. The use of RGB camera and depth sensor from Kinect to detect people is presented in [9]; however the method applies only to fixed cameras. Another similar work is the presented by Salas and Tomasi in [10], in which combine color and depth images to detect people. They use tracklets, which is a concept to refer to the same detection over a sequence and combined with Histograms-of-Oriented Gradients, the tracking of the person is performed.

III. IMAGE PROCESSING

As early mentioned, image processing is performed over monocular cameras. While no a priori knowledge is considered, i.e. dimensions or size of people or objects, it is not possible to retrieve depth information. Therefore image processing is used to detect people on a scene and then to initialize the process of 3D tracking by depth sensors.

The proposed methodology for detect people is composed of two steps, face detection and color skin detection used for validation.

A. Face Detection

The process of people detection is based mainly on the very well known method of Haar like features cascades proposed by Viola & Jones [11]. We have used the implementation proportioned by the OpenCV library. One of the advantages of this implementation is the possibility to find more than one face per image (Fig. 1.a).

Nevertheless this implementation has some deficiencies; for example in real and complex environments this method provides some false positives (fig 1.b), and besides to these results and depending of the parameters it can be a slow process (2–5Hz). Figure 1 shows the result of applying the face detector of Viola & Jones, from OpenCV library on two different scenes.

Therefore, in a second step of the proposed methodology, detected faces are filtered by merging the results of this detector with a skin color detection technique, achieving better accuracy in the people detection.

B. Skin Color Detection

Because is not the main objective of this work, and only a way to increase people detection accuracy, an analysis of different works for skin color detection has been done. In the current state of the art, many authors report the necessary data to replicate their results. Table 1 summarizes the proposed color spaces and thresholds reported on most cited papers.

<table>
<thead>
<tr>
<th>Work</th>
<th>Color Space</th>
<th>Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>YCb</td>
<td>136 &lt; G &lt; 156</td>
</tr>
<tr>
<td></td>
<td></td>
<td>110 &lt; Cb &lt; 123</td>
</tr>
<tr>
<td>[13]</td>
<td>RGB</td>
<td>R &gt; 95 &amp; G &gt; 40 &amp; B &gt; 20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&amp; max{R, G, B} = min{R, G, B}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 &gt;</td>
</tr>
<tr>
<td>[14]</td>
<td>YCb</td>
<td>Y &gt; 80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>135 &lt; G &lt; 180</td>
</tr>
<tr>
<td></td>
<td></td>
<td>85 &lt; Cb &lt; 135</td>
</tr>
<tr>
<td>[15]</td>
<td>HSV</td>
<td>0 &lt; H &lt; 18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S &gt; 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B &gt; 80</td>
</tr>
</tbody>
</table>

As color image processing techniques depends on many parameters, i.e. sensitivity of the camera or illuminations conditions, a comparative study under the common use conditions has been done. In order to select the set of parameters and color space more convenient to camera to be used, a comparison with the ground truth (manually extracted) has been done. Figure 2 show images under different conditions used to evaluate skin color detectors.
On Table 2 are shown the result of this comparative work. As can be seen, the methods given [13] and [14] provided better results. We have chosen to use the parameters given in [14] since the segmentation is faster, this is mainly due to the number of comparisons done necessary for the method reported in [13].

<table>
<thead>
<tr>
<th>Work</th>
<th>Space Color</th>
<th>Accuracy</th>
<th>% Skin</th>
<th>% No Skin</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>YCrCb</td>
<td>0.927</td>
<td>0.434</td>
<td>0.939</td>
</tr>
<tr>
<td>[13]</td>
<td>RGB</td>
<td>0.949</td>
<td>0.517</td>
<td>0.960</td>
</tr>
<tr>
<td>[14]</td>
<td>YCrCb</td>
<td>0.949</td>
<td>0.440</td>
<td>0.963</td>
</tr>
<tr>
<td>[15]</td>
<td>HSV</td>
<td>0.964</td>
<td>0.190</td>
<td>0.986</td>
</tr>
</tbody>
</table>

### C. Face Detection Merged with Skin Detection

As it has been mentioned on previous sections, returned results by face detector, including false positives, are validated by the color skin detector selected. The way as this information is merged is detailed here.

An important issue to take into account at this stage is the requirement that tracking process should work on real time, so, people detection stage should be as faster as possible to avoid losing targets when the tracking process begins.

Thus, in order to accomplish mentioned requirements, we propose a concatenation of process more than an independent parallel processing with the respective image merging procedure.

In the proposed methodology every region found by the Viola & Jones face detector is passed to the skin color detection method. If the later finds more colored pixels than a given threshold (relative to the size of the bounding box), and the compactness of the result, the region is validated as face. Figure 3 shows the result of this procedure.

### IV. RANGE DATA PROCESSING

Range sensors provide 3D data estimation that can facilitate processes of people or object segmentation. If 2D color information is merged with 3D information, the result could be more favorable that using only one type of information. This is because it can overcome the limitations that arise in the use of both types of information separately.

The complete point cloud from the scene is obtained in real time with the Kinect sensor using PCL library, however most of the algorithms proposed on this library does not work on real time. In order to deal with this constraint, we have implemented most of them under specific conditions.

The proposed methodology works as follow: once faces are detected by the previous stage, spatial position (3D) is used to detect the body of the person by a region-growing algorithm under the 3D data. A color histogram of the person is the constructed in order to follow it under next frame of the sequence avoiding with other persons or similar structures. In the following each step of this process are detailed.

#### A. 3D Region Growing

The 3D region growing technique seeks to obtain a region united by their distance, in a point cloud. This growth begins with a seed, and ends until the neighboring points do not meet the required distance values, using a threshold value of 10 cm between points.

The point seed employee for the 3D region growing of this work is the geometric center of the face, which was previously known with certainty that is a person. From this point in the image, is obtained its equivalent 3D point in the point clouds. Because it is not necessary to obtain the complete silhouette of the person, the growth stops when it reaches the region around the knees (a given altitude threshold). In this way, it eliminates the problem of growth toward the floor. Figure 4 shows the result of 3D region growing. As we can see, the output is the points belonging to the person.

While data is obtained by the PCL driver two problems are present. The first are little hollows at the person’s region, because point cloud from the Kinect presents sometimes data inconsistencies. The second problem are small regions of the background are obtained with the silhouette of the person. This is mainly due to the lack of a fine shutter synchronization procedure between the VGA camera and the infrared sensor. Is not possible to guarantee obtaining color data and depth data from the sensor always at the same time.
We have detected, using the ground data manually extracted, that from all the points segmented, 10% approximately correspond to noise (color). That has to be taken into account in order to avoid affecting system performance. So in order to follow robustly the persons in the presence of this noise an adaptive color histogram of the points is constructed by the merging of the last five observations.

At it has been mentioned, it is very important that a mobile robot can track people in a robust and efficient way. However, as it has been showed that robust people detection process is not always possible to be implemented in real time. Moreover, the following process of 3D region growing requires of a seed inside the region of the body, and if is not truly seeded a region growing can derive to other objects.

In order to do people tracking, we propose to use a similar approach to the presented in [16] with the main difference that it will be applied to a projection of the 3D data under a plane parallel to the floor and not directly to RGB images.

### B. Getting the Plan-View

The first step to track people from color histograms is to remove undesired 3D information as for example the data from floor and ceiling. By a floor plane detection method, not detailed on this paper, we recover spatial information about the main plane corresponding to the floor. From plane equation founded, cutting data between 20cm and 2m above the floor plane reduces 3D point cloud.

The resulting 3D point cloud is projected using plan-view approach. In this way, a transformation of the original coordinates is realized: the $Z$ coordinate in 3D space becomes $X$ in the density map, the $X$ coordinate in 3D space becomes $Y$ in the density map, and the $Y$ coordinate in 3D space is removed.

Figure 5 shows a point cloud without segmentation while Figure 6 shows the projection obtained by projecting the plan-view analysis.

### C. People tracking using color histograms

The 3D region growing returns, based on the points and histogram that correspond to a person, the position of this person in the projection of the projection plane. These points are enclosed in a square box, for tracking, using a particle filter with color histograms.

The set of particles is generated using a Gaussian distribution approach under the previous position of the person, in the projection plane. After obtaining the particles, the points belonging to each particle are obtained, and these will generate their histograms. Histogram from the particle that has a smaller difference with the histogram of the person, i.e., less weight, will be the position where the person is in the actual frame.

### D. Inclusion of people in the scene

Because new people can enter in the scene, the detection procedure is activated at a given frequency so every certain number of frames. Thus, the process is not affected in their performance. If the module finds people who are not being tracked, it generates a new tracking target, in addition to those previously generated.
V. EXPERIMENTAL RESULTS

In order to test the performance and efficiency of the algorithm, we have run the module with several data sequences. The process of detection and tracking was observed through the displaying of the RGB image and the projection plane from the analysis plan-view, validating the region tracked with the person’s trajectory.

In Figure 7 it is showed the results of applying our system in some frames, with their respective projection plane. The red square box represents the area where is the person. Figure 8 shows the result with another sequence, tracking one person. Figure 9 shows the result with another sequence, detecting and tracking more than one person. With the information from the Kinect, and the position from the tracking, the trajectory of the person can be recorded, for further analysis.
Because the first phase of people detection is through the face, the person must be facing the sensor; just one or two seconds. The module runs online at 15 Hz, approximately.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented a module for people detection and tracking, in real time, at this stage we have used the Kinect sensor, however as is has been said, it can be applied to other similar RGB-range systems. This module has been developed to improve Human-Robot Interaction, detecting real persons, and performing the tracking of its trajectories. Merging information from multiple sources, the module is able to obtain more accurate results than if we just use information from one source.

2D space is used to detect characteristics of people, like face and skin, while 3D is used to segment the people silhouette, because the distance information makes the segmentation more effective. Both process runs independently at difference frequencies to share information.

The condition to be mounted on a mobile robot, and to works in real time has been achieved, for that robot must be capable of self-locating by an SLAM method. The person does not need to wear clothing or accessories specials for the detection nor tracking, since the tracking is performed with the color histogram constructed from the 3D person silhouette.

In future work, some improvements to the module are:

- Add other features for detection of the person, so increasing the robustness of the initialization.
- Create a torso model or whole human body model for gesture recognition.
- Increase the ability to recovery people tracking, due to occlusion, or leaving the scene momentarily.
- Analyze the recorded trajectory using the work [17]
- Test with other sensors.

REFERENCES