Toward Real-time Human Detection and Tracking in Diverse Environments

Nathan Koenig
iRobot Corporation
63 South Ave
Burlington, MA 01803 USA

***Abstract*— Human-robot interaction (HRI) encompasses numerous disciplines from psychology to mechanical engineering. Each field contributes to a robot's ability to better understand and respond to humans actions, and behaviors. A common thread is the need to detect a person and their location within an environment. In this paper we tackle the issue of identifying which objects in a scene are people. We also discuss current work on human detection, and strategies for improved performance.**

I. Introduction

Detection of humans within a local environment is critical for HRI. Identification of individuals, gesture recognition, and understanding group dynamics are just a few of the capabilities that rely on human detection. The goal of our work is to move toward a more reliable and robust system, where human detection is viewed as an fundamental stepping stone toward rich HRI rather than an end goal.

The process of achieving this goal is non-trivial. Humans are generally hard to quantify in a clean manner. People come in all different shapes, sizes, and color. We wear vastly different clothing, move frequently and sometimes rapidly, congregate in large and small numbers, and interact in a world with varied lighting conditions and clutter. There are few features easily identified among all humans. These factors make it difficult to generate general purpose robust algorithms to detect humans.

Another complexity arises from the wide range of sensors used for human detection. These can include mono and stereo cameras, laser range finders, sonar, and infra-red cameras. A range of algorithms can also be applied to these sensors, such as face detection, motion analysis, model-based and learning methods. The result is a wide range of techniques and strategies for human detection and tracking that work given a set of conditions on sensing devices, algorithms, and environmental conditions.

In this paper we present a fast and robust approach to human detection at short ranges. As with most approaches, ours has a few caveats. The sensor we use is the Swiss Ranger SR-3000 3D camera, see Figure 1. The properties of this camera limit the sensing range and field of view to fairly narrow proportions. Given this factor, we have chosen to use a connected components algorithm that leverages the high density of data the Swiss Ranger returns at short distances.

Based on current experience with the Swiss Ranger and past experience with lasers and stereo vision, we propose that a single sensor-algorithm combination is not an ideal solution for human detection. Given that humans rely on sight, sound, smell, and touch to identify objects it seems inappropriate to limit a robot to just one sensor and algorithm. Some will argue that vision, a rich source of feature information, can and will solve most robot sensing issues. To a large extent this is true, but current technology is not at this level. In the mean time we should not limit advances in other aspects of HRI. A solution involving multiple sensors working in conjunction is an attractive solution to provide robust and long-term human detection and tracking.

II. Related Work

Significant work has used motion analysis to detect humans [2], [5], [7], [12], [15]. These approaches are heavily reliant on custom models, and work with fixed cameras. A camera mounted on a mobile robot will encounter difficulties separating human motion from robot motion.

Laser based approaches use scanning laser range finders to detect and track humans [4], [11]. This technique is simple to implement, requires little processing power, and allows for fast tracking. However, detection is limited to a single plane. Numerous objects such as trees and poles can be falsely detected, and tracking multiple people who cross paths is difficult.

Model based methods use the structure of the human body to detect people in images [14], [16]. These approaches find parts of the human body, such as the head, hands, feet, and legs in a scene and construct a relationship between each component. Positive detections occur when the parts and their relationships match a model. This method is prone to error with cluttered environments, partially obscured people, and unexpected types of clothing such has hats, gloves, long coats, and dresses.

A promising technique of human detection relies on Haar wavelets [9] used in conjunction with an support vector machine (SVM) [10], [13]. Haar wavelets encode relationships between neighboring regions in an image. These wavelets are used to build a *wavelet template*, that captures the structural relationship between instances of a class at various scales. The templates are then used as the feature vector to train an SVM. This approach has a detection rate of 70% with a false positive rate of 1:15,000. Their results are based on images of pedestrians in cluttered outdoor environments.

Inspired by SIFT [8] descriptor, histograms of oriented gradients have been developed to accurately detect humans in

Fig. 1. Swiss Ranger SR-3000 3D time-of-flight camera.

arbitrary scenes [3]. This method divides an image into densely overlapping cells, where each cell consists of a histogram of oriented intensity gradients. These features are then used to train and SVM. Performance of this method is significantly better than wavelet based approaches, however processing each frame is computational expensive.

III. System overview

The goal of our system is to detect, track, and follow a human in real-time over large time-spans in arbitrary environments. Our prototype system consists of a Swiss Ranger 3D camera, see Figure 1, mounted on an iRobot Create mobile robot, see Figure 3. The Swiss Ranger acts as the primary sensor, and generates a dense 3D point cloud. The Create mobile robot is used as a development platform, and has served well for rapid prototyping. Future work will use an iRobot PackBot, see Figure 2 as the mobile base, thereby increase the range of applicable environments.

A connected components algorithm processes the data generated by the Swiss Ranger to produce a set of large solid objects within the camera's field of view. Following this step, an SVM identifies which of the objects are human. The SVM has been trained on the shape of a persons profile as seen by the Swiss Ranger after the connected components has completed. The final step of tracking the person is achieved using a PD controller.

Fig. 2. iRobot PackBot EOD mobile robot.

IV. Swiss Ranger: 3D sensor

The Swiss Ranger camera consists of an array of near-infrared light emitting diodes and a custom CCD image sensor. All the LEDs flash modulated infrared light. The CCD sensor, 176x144 pixels in size, detects the reflected infrared light. Internal computation measures the time-of-flight for each light beam detected by the CCD's pixels. Data returned by the Swiss Ranger consists of a 3D vector and intensity value for each pixel in the CCD. Figures 4(a) and 4(b) demonstrate a sample point cloud. Figure 4(c) depicts the corresponding intensity data.

Fig. 3. iRobot Create mobile robot, with Swiss Ranger camera.

The field of view of the Swiss Ranger camera is 47.5 x 39.6 degrees, with a maximum range of 7.5 meters. Data is generated at a rate dependent on the range of the reflector, see Table I. Since we are detecting and tracking humans, the Swiss Ranger is placed at height of roughly three feet, see Figure 3. This height provides an unskewed view a person, and allows the camera to easily see a person's head and torso.

While the Swiss Ranger is convenient to use, it does have significant limitations. The short field of view and range make tracking a human difficult. A person moving even at a normal walk quickly leaves the camera's field of view. This in turn requires the robot to rotate significantly, introducing noise.

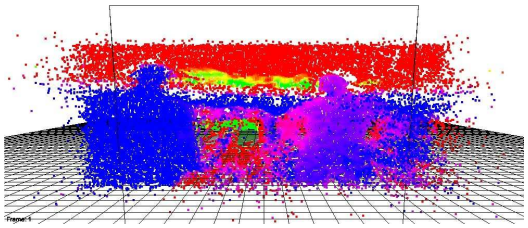
The stated range of the camera is 7.5 meters, although it's usable range is limited to at most five meters. At greater distances, objects tend to blur together in terms of both their intensity and depth values. Objects closer than one meter are too large to identify as human, and cause significant measurement noise. The usable range of the camera is therefore limited from one to five meters.

Highly reflective objects, such as shiny metal surfaces, generate false readings. The camera reports max range readings in these cases. Such reflective objects can also skew surrounding readings, depending on distance to the camera. The closer a highly reflective object, the greater the affected region. This is based on empirical data gathered from use in a typical office environment. In our case, highly reflective objects mainly included metal surrounding light fixtures, and occasionally a person's eye. Figure 4(c) contains examples of these errors in the ceiling lights fixtures, which are marked black.

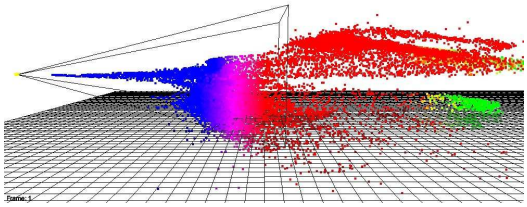
Distance(m)	0.3	1	2	3
Frame Rate(Hz)	29	20	15	12

TABLE I

Swiss Ranger distance and frame rate table.

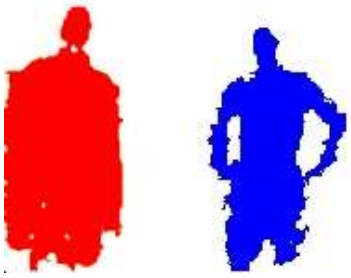


(a) Point cloud front view



(b) Point cloud side view

(c) Intensity image



(d) Connected components

Fig. 4. Sample data returned from Swiss Ranger camera.

V. Mobile Robot and Computation

We have used the iRobot Create mobile robot as a development platform. This recently released and inexpensive robot is capable of basic movement, is equipped with a bump sensor, and provides odometry data. External computation can be connected to the robot via a 25-pin UART header, which provides access to power, state information, and motor control.

The Create robot has allowed us to rapidly prototype our human detection and tracking system. Very little setup is required to use the Create, and extensive documentation is provided on the robot's API [1], and how to interface other hardware [6]. This robot is limited in terms of it's speed and scalability. Future work will replace the Create with an iRobot PackBot.

A PackBot is a robust platform that provides all-weather and all-terrain mobility. This type of robot is well suited for tracking people in a variety of environments. More computation and sensors can be carried and powered by the PackBot. Currently, the Create robot is limited to flat indoor environments with slow moving people. A PackBot on the other hand will be capable of tracking fast and slow people both indoors and out, and up and down stairs.

Both the Create and PackBot robots have insufficient onboard computation for this vision based task. In order to gather and process the Swiss Ranger data in timely manner we have used a modern laptop. This laptop consists of a Pentium Core 2 Duo 2.0GHz processor with one gigabyte of ram. With this hardware, the algorithms presented run at approximately 15 Hz. This frame rate is fast enough to track a person moving at a moderate pace, but insufficient for brisk walks and runs. Future software iterations will attempt to push the frame rate to 30 Hz.

VI. Human Detection

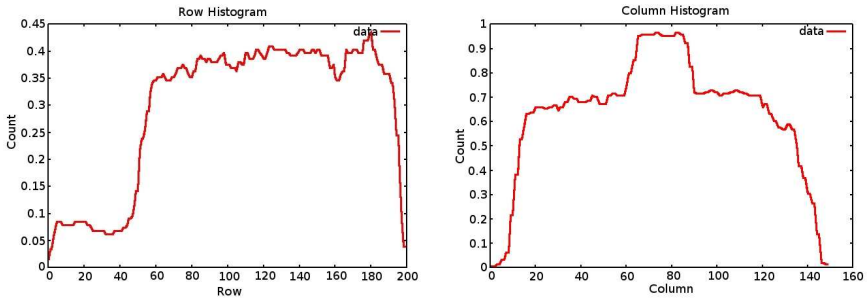
Our detection algorithm consists of two functions. The first is a routine to find all objects in a scene that could potentially be human. The second function identifies which of the objects are human. Tracking of the people is then accomplished using a Kalman filter, and a PD controller allows the robot to follow a moving person.

The first phase of human detection, where all potentially human objects are found, relies on the observation that contiguous objects have slowly varying depth. In other words, a solid object has roughly the same depth, or Z-value in our case, over its visible surface. An algorithm capable of detecting these solid surfaces is ideally suited for finding these objects.

We have chosen to use a connected components algorithm based on its speed and robustness. This algorithm groups together pixels in an image based on a distance metric. For our purposes, each pixel is a point in 3D space, and the distance metric is the Euclidean distance along the Z-axis between two points. When the distance between two points is less than a threshold value, the two points are considered to be part of the same object. The output of the algorithm is a set of groups where each group is a disjoint collection of all the points in the image. A simple heuristic of eliminating small connected components, e.g. those with few points, significantly reduces the number of components. The final result is depicted in Figure 4(d).

The second phase of our human detection algorithm identifies which of the remaining connected components are human. To solve this problem we have trained an SVM on the shape of a human. The trained SVM is then used to identify which of the connected components are human and which are not.

An SVM is a learning algorithm used in pattern classification and regression. The working principal behind an



(a) Row histogram (b) Column histogram

Fig. 5. Row and column histogram of right component in Figure 4(d).

SVM is to project feature vectors into a higher order space where separating hyperplanes can classify the data. Our feature vector consists of the shape of the human in the form of a row-oriented and column-oriented histogram. For a given connected component, the row-oriented histogram is computed by summing the number of points in each row of the connected component. The column-oriented histogram is computed based on data in the columns of the connected component. Figures 5(a) and 5(b) depict the row histogram and column histogram from the connected component found on the right in Figure 4(d). Before computing the histograms, the components are normalized to a constant size of 200x160 pixels.

Our SVM approach to human detection is still in the preliminary stages of development, however results are promising. Using an SVM to identify humans from a set of features is not new [3], [9], [10], [13]. These prior uses all have carefully chosen their feature set in order create a large separation between human and non-human. In a similar manner, the dual histograms of the human shape is fairly unique due to the shape of the head and shoulders.

Our technique for human detection will likely be robust to a wide variety of people, clothing, lighting conditions, and to a lesser extent clutter in the environment. Detection relies only on a nearby, one to five meter range, person whose head, shoulders, and torso are visible. This method does degrade rapidly when the person is partially occluded.

VII. Tracking and Following

Tracking of the people is accomplish via a Kalman filter, which estimates the future pose of a person, and then corrects based on observations. A Kalman filter's update cycle is fast, and has seen wide spread use in real-time systems. This approach provides an efficient means to follow a single moving object, in this case a human, in the presence of uncertainty.

The final component is motor control to follow a detected human. This is accomplish using a PD, proportional and derivative, controller based on the observed pose of the human. This type of controller has a fast update cycle, and has proved capable of following the movements of a human. Future work will incorporate local and global path planning to avoid obstacles and generate efficient trajectories based on the kinematics of the robot.

VIII. Discussion

The sensor we have used is reflected in our algorithmic approach. A Swiss Ranger camera provides dense 3D data at close ranges. It lacks color information, and a wide field of view. As a result we rely heavily of the 3D data to pick out connected components in the scene.

The results we have received to date show that this approach will work well when few people are in the scene and they are not partially occluded. If these conditions are met, then this approach can segment the camera data and detect humans at roughly 15Hz. However, these preconditions are not very likely, especially in highly dynamic environments, and therefore we must explore alternative approaches.

With access to other sensors, the most popular being a monocular camera, numerous other methods are available. The most promising approach uses a dense grid of histograms of oriented gradients to extract features from and image. Experiments using this method achieve greater than 80% accuracy when processing scenes with multiple people in cluttered environments and various lighting conditions. The primary detractor to this method is it's slow frame rate.

The methods that have been described all have various pros and cons. Three dimensional data from a Swiss Ranger allows for fast segmentation algorithms, at the cost of a narrow field of view and small range. Histogram of oriented gradients can operate at long ranges in a wide variety of environments, but requires intense computation. Laser based approaches restrict detection to a single plane, can be confused with numerous people, but is fast and efficient.

We believe that a hierarchical approach is necessary for human detection. One such approach would combine a monocular camera using histogram of oriented gradients and a the Swiss Ranger running connected components. Detection of people in the periphery and at long range would be tasked to the monocular camera running in a slow update loop. Nearby people, who are more relevant for HRI, would be detected and tracked using the Swiss Ranger, or possibly a stereo camera, and connected components running in a fast update loop.

Such a combined method has the potential to leverage the benefits of both systems, and minimize their weaknesses. While this won't be a perfect system, it will hopefully combine different methods of sensing and processing in a manner that

result in a net gain in terms of performance. The goal of this work is to move closer towards a simple and robust system for human detection. This goal is similar to how robot localization has progressed into a well defined solution that provides a base upon which numerous other robot tasks are built. Human robot interaction relies on human detection, and progress in this field can greatly benefit from a reliable and robust system.

REFERENCES

[1] irobot create open interface. iRobot Corporation. `http://www.irobot.com/filelibrary/create/Create Open Interface_v2.pdf`.

[2] H. J. Chen and Y. Shirai. Detecting multiple image motions by exploiting temporal coherence of apparent motion. In *In proceedings of Computer Vision and Pattern Recognition*, 1994.

[3] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, 2006.

[4] Ajo Fod, Andrew Howard, and Maja J. Matarić. A laser-based people tracker. In *IEEE International Conference on Robotics and Automation*, pages 3024–3029, Washington DC, May 2002.

[5] B. Heisele and C. Wohler. Motion-based recognition of pedestrians. pages Vol II: 1325–1330, 1998.

[6] Nathan Koenig, David Feil-Seifer, and Maja Matarić. Robotics primer workbook. Interaction Lab, University of Southern California. `http://roboticsprimer.sourceforge.net`.

[7] M.S. Lee. Detecting people in cluttered indoor scenes. pages I: 804–809, 2000.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. 60(2):91–110, 2004.

[9] S. Mallat. A theory for multiresolution signal decomposition: The wavelet represenstation. 11(7):674–693, July 1989.

[10] Michael Oren, Constantine Papageorigiou, Pawan Sinha, edgar Osuna, and Tomaso Poggio. Pedestrian detection using wavelet templates.

[11] Anand Panangadan, Maja J. Matarić, and Gaurav S. Sukhatme. Detecting anomalous human interactions using laser range-finders. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2136–2141. IEEE Press, Sep 2004.

[12] K. Rohr. Incremental recognition of predestrians from image sequences. In *In proceedings of Computer Vision and Pattern Recognition*, 1993.

[13] Hiroaki Shimizu and Tomaso Piggio. Direction estimation of pedestrian from multiple still images.

[14] N. Sprague and J. Luo. Clothed people detection in still images. pages III: 585–589, 2002.

[15] Paul Viloa, Michael J. Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appereance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003)*, Nice, France, 2003.

[16] L. Zhao. *Dressed Human Modeling, Detections, and Parts Localization*. PhD thesis, Carnegie Melon University, Pittsburgh, PA, 2001.