

A FORMAL EXPERIMENT TO ASSESS PEDESTRIAN DETECTION AND TRACKING TECHNOLOGY FOR UNMANNED GROUND SYSTEMS

Barry A. Bodt*
U.S. Army Research Laboratory
APG, MD 21005

ABSTRACT

An important area of investigation in robotics perception and intelligent control concerns the ability to detect, track, and avoid humans operating in proximity to an unmanned ground vehicle (UGV). Under the Army Research Laboratory (ARL) Robotics Collaborative Technology Alliance (RCTA), ARL and other member organizations have developed algorithms focused on human detection and tracking, which leverage program advances in stereovision and LADAR. A recent assessment conducted by ARL and the National Institute of Standards and Technology (NIST) exercised these technologies under relevant conditions. This paper highlights technology advances demonstrated in this investigation. The most significant findings are that pedestrians can be reliably detected and tracked and that with the inclusion of temporal filtering on algorithm reports, incidences of misclassification of other objects as pedestrians can be dramatically reduced.

1. INTRODUCTION

An important area of investigation in robotics perception and intelligent control concerns the ability to detect, track, and avoid humans operating in proximity to an unmanned ground vehicle (UGV). Under the Army Research Laboratory (ARL) Robotics Collaborative Technology Alliance (CTA), ARL and other member organizations have developed algorithms focused on human detection and tracking, which leverage program advances in stereovision and LADAR.

This work is the third in a series of investigations. Camden and Bodt (2006) reported that 98 of 101 stationary, upright mannequins (human surrogates) were detected as humans during autonomous operation of the ARL Experimental Unmanned Vehicle (XUV) relying on LADAR for perception. Barrels were misclassified as humans 58% of the time. Platform speeds in this study never exceeded 15 kph and MOUT conditions were not considered. Rigas et al. (2007) detailed a more thorough investigation, building on the previous study. Clutter consistent with a MOUT environment was included along the course, XUV speeds were increased to a maximum of 30 kph, some mannequins were moving and

in different postures, moving target vehicles were added, and detection reports were achieved for three algorithms simultaneously. (The XUV was tele-operated in this study to ensure safety and to provide a view for all algorithms uninfluenced by the autonomous navigation system.) In this more complex exercise, algorithms detected moving mannequins in excess of 80% of the time, and fixed mannequins in excess of 60% of the time. A limitation of the study, however, was that ground truth for moving mannequins mounted on a rail system was difficult to achieve.

In September 2007 a third experiment was conducted. The paper reports on this third study, details improvements in the experimental approach consistent with three principal objectives, and reports new results for pedestrian detection and tracking.

2. EXPERIMENTAL APPROACH

The present investigation balances multiple objectives. The overarching goal was to expose the algorithms and sensors on board an operated Suburban to complex pedestrian traffic using human subjects and to observe algorithm performance in detection and tracking. A secondary goal was to explore the impact of relevant conditions (e.g., platform speed, pedestrian speed, MOUT conditions). A tertiary objective, important to program participants, was to advance the experimental methodology to yield greater information in the feedback loop to developers. We address each of these in turn.

2.1 Human Detection

This assessment marked the first time in this program that human movers acted as targets for detection from a moving vehicle. Events include humans advancing and retreating from the vehicle at different angles, humans crossing paths in close proximity and occlusion situations where sight to the mover from the sensor system is momentarily lost. Repeatable human movement scenarios relative to the movement of the vehicle were choreographed to ensure a consistent presentation of the complex event to the sensor systems. Ten pedestrians were used in each run. Figure 1 illustrates the paths of 7 humans relative to the path of the Suburban. The remaining three

humans followed random chords within the open circle. The data supports comparative analysis across treatment conditions and allows developers to examine performance with respect to detection events.



Fig. 1 Human paths (dashed line), mannequin locations (solid circles), Suburban path (solid line), and random human motion (open circle) on the test course.

2.2 Relevant Conditions

A secondary objective was to explore the impact of relevant conditions. Pedestrian scenarios were replicated in accordance with an experimental design incorporating terrain (MOU and open), vehicle speed (15 and 30 kph), and pedestrian speed (1.5 and 3.0 m/s) over 32 runs. The 250 m test course included some clutter from natural vegetation along with numerous man made obstacles (e.g., fire hydrants, barrels, and posts). Figures 2 and 3 picture detection events on one run. Algorithms reported human detections at data frame rates ranging from 2.69 to 18.3 Hz based on a broadcast sensor frame rate of 10 Hz. Response measures included the probability of detection, probability of misclassification (other obstacles reported as humans), the number of false alarms (no known obstacle), as well as measures to quantify continuity and persistence of tracking.



Fig. 2 Suburban equipped with sensors and algorithm processors (shuttles) passes a truck and jogging humans.

2.3 Improved Methodology

A tertiary objective, important to program participants, was to advance the experimental

methodology to yield greater information in the feedback loop to developers. In keeping with that goal, algorithms were used simultaneously during a run by allocating individual computer shuttles to each for processing and by distributing the sensor information at higher frame rates. This allowed direct comparison of algorithms within a run. In addition, time-stamped ground truth, difficult for real-time pedestrian traffic, was accomplished with the introduction of an ultra wideband (UWB) wireless tracking system implemented by NIST. This system provided precise time and location for pedestrians that could be compared with algorithm reports. See figure 4.



Fig. 3 Suburban passes obstacle clutter and encounters human movers with crossing tracks.

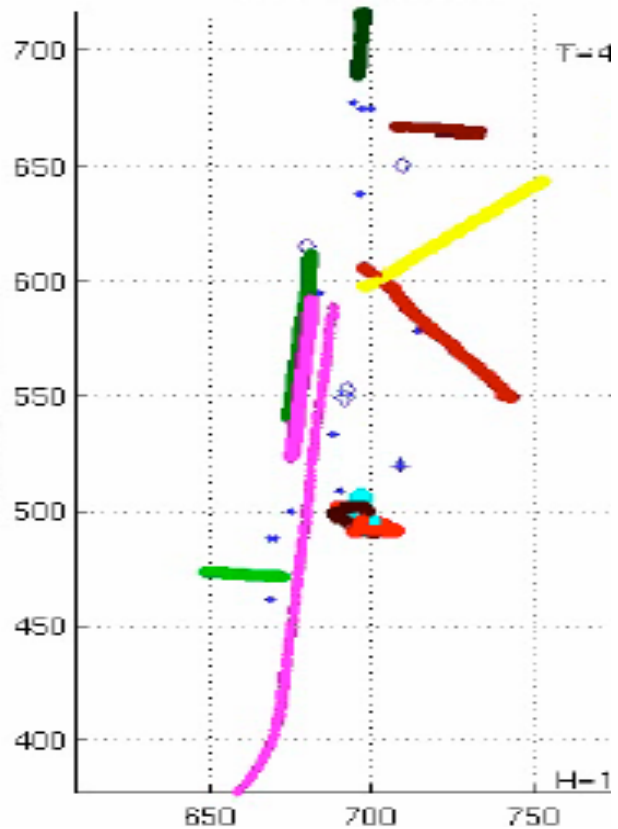


Fig. 4 UWB wireless tracks of humans and Suburban located by easting (x) and northing (y) during one run.

3. DESIGN AND ANALYSIS

In this section we offer an overview of the experimental design (e.g., sources of data, manner of collection) and the analysis implemented.

Seven algorithms yielded data during the study. Participating RCTA members included Carnegie Mellon University (CMU), General Dynamics Robotics Research (GDRS), ARL, the Jet Propulsion Laboratory (JPL), and the University of Maryland (UMD). Five algorithms were based on LADAR (CMU [2], GDRS [2], and ARL) and two were based on stereovision (JPL and UMCP). CMU1 was a SICK LADAR. CMU2 was a 3D LADAR reduced to SICK. Rigas et al. (2007) lists details for how detection was accomplished for each algorithm.

The site used, shown in figure 1, was a section of improved public road, not yet open for public use, located behind General Dynamics Robotics Research in Westminster MD.

The experimental design was conducted as a three-factor factorial design with four replications over 32 runs. The factors of the study were human and vehicle speeds and terrain type, MOUT or open. A randomized run schedule was developed and strictly followed. Frequent calibration of the UWB wireless was interspersed in the run schedule to ensure accuracy of ground truth. Choreography of human paths relative to the suburban track was carefully administered to ensure that under varying experimental conditions the sensor perspective to all complex events was the same across runs. Test protocol included controls to ensure data had been captured prior to proceeding to the next run.

Analysis began with post processing of the sensor data to align with ground truth objects and humans. A detection called by the algorithm signified that a human was present at that location. All algorithm detections were compared with ground truth. Detections within 5 m of a human ground truth were valid detections. Detections within 5 m of another object type were considered misclassifications and detections further than 5 m from any known ground truth were labeled false positives.

Data analysis initially focused on summary statistics and graphical analysis pertaining to the probabilities of detection and misclassification, along with the frequency of false positives. This analysis was augmented with video and Matlab movies comparing the algorithm outputs to the ground truth for each run. The impact of design factors was addressed with analysis of variance. During this analysis, a temporal filter was imposed on the algorithm reports. Developers had been instructed to

report detections each frame. But this approach led to a large percentage of misclassifications. We explored the impact of requiring that detections be persistently tracked for at least a few frames, rather than simply reporting an instantaneous finding by each algorithm for each frame.

4. RESULTS

Results are reported consistent with the three objectives of the study: human detection, relevant conditions, and improved methodology.

4.1 Human Detection

We begin with the simple listing of the percentage detection, percentage misclassification and the number of false positives recorded for each algorithm based on as little as one frame of data. Those results appear as Table 1.

Focusing on the percentage of detections, we see very good performance for all algorithms except CMU2. We should note that there were known calibration issues with that algorithm. The vision systems report is based on only seven of the ten humans on the course. A more limited field of view placed almost all of the human movement within the circle outside the sensor range. Two other choreographed human tracks were just within the sensor range. Almost all remaining missed detections for the vision systems over the 32 runs were from those two humans.

Table 1. Summary Algorithm Performance

Algorithm	% Detect	% Misclassify	# False Positives
ARL1	99.6	75.8	1522
CMU1	94.1	1.5	171
CMU2	31.9	1.5	4
GDP1	99.4	35.0	460
GDW1	100.0	56.7	1590
JPL1	87.9	22.5	55
UMD1	89.3	20.6	76

Dynamic planning will ultimately benefit from correct classification as well as detection. Misclassifications occurred at low rates for CMU1, even with a high percentage of detection and low numbers of false positives. GDW1 and ARL1 showed the greatest number of misclassifications, initially, in addition to a high number of false positives.

During the analysis, it became clear that results based on a single data frame were different than an algorithm determination based on a few to several frames. Further analysis was performed in which a temporal filter was imposed ensuring at least two contiguous data frames to at

least ten data frames upon which the algorithm detection decision would be based. (Filtering was not possible for ARL1 because reported data did not support tracking). Table 2 shows results for three or more data frames of persistent tracking. Note the large reduction in the percentage misclassification and the number of false positives achieved by this adjustment. For example, GDP1 gave up just 3.1% in detection but cut its misclassification percentage to ~ 25% of its original value, while the number of false positives were reduced to ~ 40% of the original value. Table 3 shows results for five or more data frames of persistent tracking. We see from examination of this table that additional gains in the tradeoff between detections and misclassifications and false alarms are not as great as when the filter was imposed for at least three data frames of persistent tracking.

Table 2. Summary Algorithm Performance (Three or More Frames of Persistent Tracking)

Algorithm	% Detect	% Misclassify	# False Positives
ARL1	-	-	-
CMU1	90.6	1.3	129
CMU2	22.2	0.5	0
GDP1	96.3	9.9	181
GDW1	100.0	18.3	800
JPL1	85.3	16.3	27
UMD1	86.6	12.7	46

Table 3. Summary Algorithm Performance (Five or More Frames of Persistent Tracking)

Algorithm	% Detect	% Misclassify	# False Positives
ARL1	-	-	-
CMU1	86.6	1.2	99
CMU2	15.6	0	0
GDP1	92.8	6.3	121
GDW1	100.0	15.5	610
JPL1	74.6	10.7	18
UMD1	67.9	6.9	24

False alarms reported may be overstated. A false alarm is called when the detection location reported by the algorithm does not agree with a known ground truth to within 5 m. However, graphical analysis in some cases suggested the detection was not spurious but rather was misclassified. For example, the 460 false alarms credited to GDP1 were all clustered in nine locations. Review of the video records revealed items (e.g., chairs for humans resting between runs, a cooler of water) that were on the course but were not recorded as known objects.

Another area of investigation concerned which object types were more likely to be misclassified as

humans. In figures 5 and 6 the results for one LADAR based algorithm (GDP1) and one vision based system (UMD) are shown. As suggested by the previous discussion, most of these misclassifications are greatly reduced or vanish altogether under temporal filtering. Still, it is useful to know which object types require greater scrutiny before making a determination. An interesting result was that large crates and trucks were often misclassified as humans. We suspect that some of this is due to human tracks coming in close proximity to the trucks and crates, together with the variability associated with the algorithms providing exact locations for the objects detected. Human detections may have been associated with an incorrect ground truth.

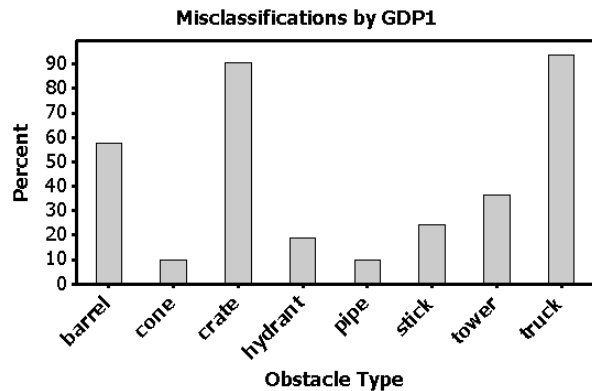


Fig. 5 GDP1 misclassifications by obstacle type with no temporal filtering.

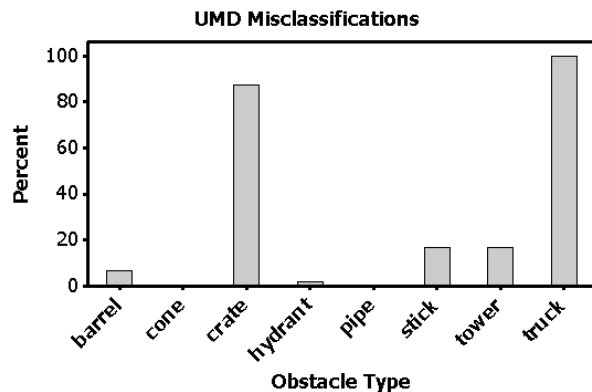


Fig. 6 UMD1 misclassifications by obstacle type with no temporal filtering.

The distance to the object at time of first detection was also noted for each algorithm and for each obstacle type over the 32 runs. Figure 7 shows this result for JPL1. The information is presented as parallel box plots based on the minimum, maximum, median, and quartiles. The box plots in green indicate humans or mannequins that should have been detected. The box plots in yellow indicate objects misclassified as humans. The median distance to first detection of humans was 27.7 m. This figure is related to the figure 8, which shows box plots of the duration of time

the objects were detected during the run. We can see from this figure that misclassified objects were often misclassified only for a short time. Then they were no longer reported as human. This effect is especially striking when viewing the results of one of the LADAR based systems, such as GDW1 shown in figures 9 and 10. Generally, misclassified objects were only reported as humans a brief duration of time.

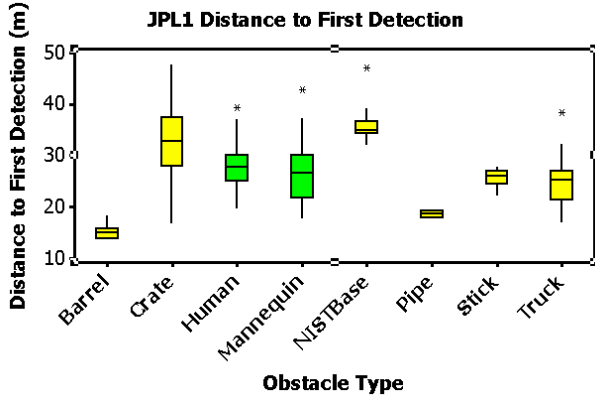


Fig. 7 Distance to first detection by obstacle type for JPL1

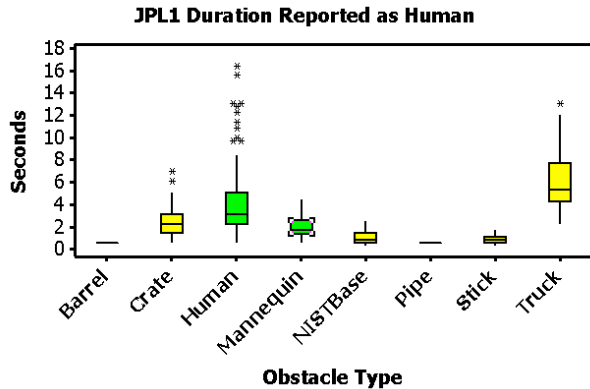


Fig. 8 Duration an obstacle type is detected as human by JPL1

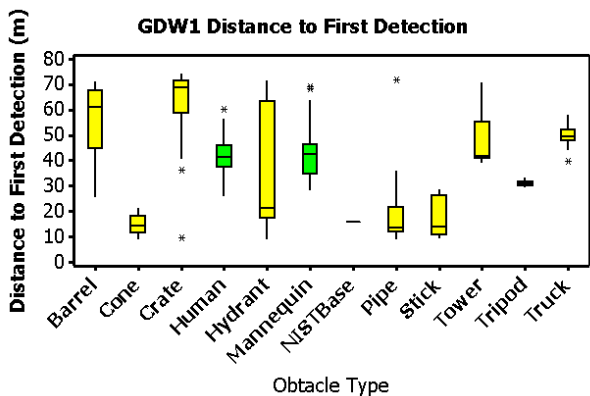


Fig. 9 Distance to first detection by obstacle type for GDW1

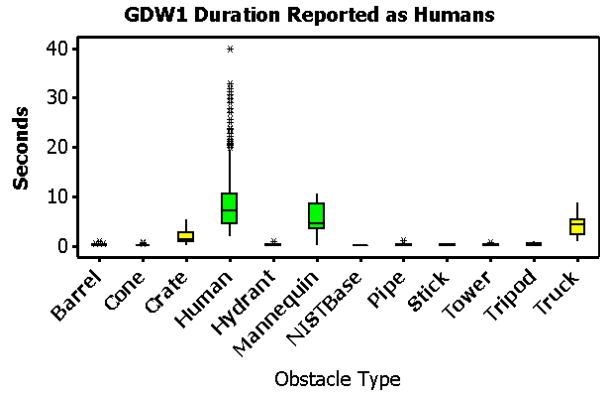


Fig. 10 Duration an obstacle type is detected as human by GDW1

A comparison of algorithms on the basis of distance to first detection appears as Table 4. The table includes the minimum, maximum, and median of the data. Note in consideration of this table that the values do not necessarily indicate sensor range, but rather when the algorithm was ready to record that a human had been detected, and this latter decision is related to the tolerance for misclassification.

Table 4. Distance to First Human Detection by Algorithm

Algorithm	Minimum	Median	Maximum
ARL	9.2	23.1	36.5
CMU1	14.0	46.1	62.8
CMU2	6.0	27.1	43.2
GDP1	18.5	28.2	37.8
GDW1	25.6	41.1	56.2
JPL1	19.6	27.7	37.0
UMD	20.9	29.5	37.8

4.2 Relevant Conditions

Data was partitioned to include only detections of actual humans, the focus of the study. In Table 5 we summarize the findings in terms of main effects for three response measures and only GDP1 as representative of our findings. The response measures are the number of humans detected, the number of unique IDs for a given human, and the distance the vehicle was from the human at the time of first detection. The second measure was intended to provide information on the ability of an algorithm to recognize the same human, but with a break in track. When algorithms detect a “new” human, a unique ID is assigned. Cell entries represent response averages for the conditions cited. The response standard error as reported by the analysis of variance appears in the bottom row. Most differences were statistically significant at the 0.05 level. Only the cell in gray was not. Whether the differences are practically significant is not addressed here. Note for GDP1, the higher number of unique IDs / human for MOUT may be due to breaks in the track as the human

momentarily disappeared behind a crate or truck and the number of humans detected are hindered by MOUT conditions. Lower vehicle speed increased the number of humans detected for GDP1 and increased the number of unique IDs for both algorithms. Distance away at first detection under open terrain matched intuition by allowing detection at greater distances when unobstructed by MOUT obstacles. Detection at greater distances for vehicle speeds of 30 kph is present but the cause is unclear. This observation holds for mannequins as well. There are clear differences in performance between algorithms.

Table 5. ANOVA Results for GDP1

Factor	Level	GDP1		
		Humans Detected	Unique IDs /Human	Distance First Detection (m)
Terrain	MOUT	8.06	1.79	28.86
	Open	8.88	1.58	30.12
Vehicle Speed (kph)	15	9.25	1.87	28.88
	30	7.69	1.50	30.10
Human Speed (m/s)	1.5	8.69	1.61	30.31
	3.0	8.25	1.76	28.67
	Overall	8.47	1.69	29.49
	Error	0.70	0.17	1.25

4.3 Methodology

The methodology introduced in this experiment advanced our experimentation capability. Moving humans provided the realism that was lacking in previous studies. The choreography of those humans ensured that each of the relevant condition factors were examined under similar perspectives to the human events unfolding on the course. Finally, the UWB wireless tracking of humans provided a flexible system for reliable ground truth. Whereas in the previous studies, ground truth was elusive, this system completely specified the locations of tracked individuals during the entire run.

CONCLUSIONS

All objectives for assessment were met and several key results were established as a result of this study. Algorithms developed under the RCTA performed admirably. The detection probability for some algorithms neared 100%, but misclassification error based on

immediate, one-frame decisions remained high (>50%) for some obstacle types. During the analysis, temporal filtering was employed to require sustained or persistent tracking of a declared human for multiple frames before accepting the algorithm's detection. This ROC curve sensitivity analysis showed that by requiring only a few frames of persistent tracking that misclassification of other obstacles as human could be greatly reduced or eliminated. The distance between the pedestrian and vehicle at time of first detection varied according to the sensor system. LADAR supported algorithms performed best with regard to distance. For example, the average distance from the pedestrian at time of first detection would conservatively support 3 seconds of planning and execution for avoidance of a predicted collision with the autonomous vehicle traveling at 30 kph. Track continuity for movers need additional work to reduce the risk of confusion in avoidance planning. Algorithms misclassified other objects as pedestrians most often when reporting detections at the limits of the sensor. As the vehicle came closer, the likelihood of misclassification was greatly reduced.

The practical significance of vehicle speed, pedestrian speed, and terrain are mixed. Results are specific to algorithms, but a few general observations can be made. Effects such as reduced detections for increased vehicle speed are intuitive. Similarly, expected observations were made involving increased detections and distance to first detection in open terrain. Algorithms do not always recognize the same mover in successive frames and so record this by assigning a unique algorithm ID. More IDs per mover are seen under MOUT conditions where temporary occlusions occur.

Based on findings from this study, developers and FCS have provided input for ranking more complex human detection challenges, notably humans presented in various postures traveling nonlinear tracks at variable speeds with more occlusion possibilities. These challenges will be explored in follow-on investigations.

REFERENCES

Camden, R. and Bodt, B., 2006: Safe Operations Experiment Report," ARL-TR-3773, U.S. Army Research Laboratory: Aberdeen Proving Ground, MD.

Rigas, E., Bodt, B., Camden, R., 2007: Detection, tracking, and avoidance of moving objects from a moving autonomous vehicle, Proc. SPIE 6561.