

Stereo-vision based perception capabilities developed during the Robotics Collaborative Technology Alliances program

Arturo Rankin*, Max Bajracharya, Andres Huertas, Andrew Howard, Baback Moghaddam, Shane Brennan, Adnan Ansar, Benyang Tang, Michael Turmon, Larry Matthies
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA, USA 91109

ABSTRACT

The Robotics Collaborative Technology Alliances (RCTA) program, which ran from 2001 to 2009, was funded by the U.S. Army Research Laboratory and managed by General Dynamics Robotic Systems. The alliance brought together a team of government, industrial, and academic institutions to address research and development required to enable the deployment of future military unmanned ground vehicle systems ranging in size from man-portables to ground combat vehicles. Under RCTA, three technology areas critical to the development of future autonomous unmanned systems were addressed: advanced perception, intelligent control architectures and tactical behaviors, and human-robot interaction. The Jet Propulsion Laboratory (JPL) participated as a member for the entire program, working four tasks in the advanced perception technology area: stereo improvements, terrain classification, pedestrian detection in dynamic environments, and long range terrain classification. Under the stereo task, significant improvements were made to the quality of stereo range data used as a front end to the other three tasks. Under the terrain classification task, a multi-cue water detector was developed that fuses cues from color, texture, and stereo range data, and three standalone water detectors were developed based on sky reflections, object reflections (such as trees), and color variation. In addition, a multi-sensor mud detector was developed that fuses cues from color stereo and polarization sensors. Under the long range terrain classification task, a classifier was implemented that uses unsupervised and self-supervised learning of traversability to extend the classification of terrain over which the vehicle drives to the far-field. Under the pedestrian detection task, stereo vision was used to identify regions-of-interest in an image, classify those regions based on shape, and track detected pedestrians in three-dimensional world coordinates. To improve the detectability of partially occluded pedestrians and reduce pedestrian false alarms, a vehicle detection algorithm was developed. This paper summarizes JPL's stereo-vision based perception contributions to the RCTA program.

Keywords: Stereo vision, passive perception, terrain classification, learning, obstacle detection, pedestrian detection, water detection, mud detection, vehicle detection

1. INTRODUCTION

On May 31, 2001, the U.S. Army Research Laboratory (ARL) commenced an eight year program called Collaborative Technology Alliances (CTA) to further technology development in areas of importance to future Army capability. The program scope covered five areas of technology: advanced decision architectures, advanced sensors, communications and networks, power and energy, and robotics. The projected scope of each CTA was approximately \$35 million for a base period of five years plus \$20 million for an optional three-year extension. General Dynamics Robotic Systems (GDRS) was the lead industrial organization for the Robotics CTA (RCTA). GDRS led a consortium of industrial, academic, and government laboratory partners that included Carnegie Mellon University (CMU), Applied Systems Intelligence, Jet Propulsion Laboratory (JPL), Alion Science and Technology, BAE Systems, Sarnoff Corporation, SRI International, Florida A&M University, University of Maryland (UMD), PercepTek, Robotic Research, Signal Systems Corporation, Howard University, North Carolina A&T University, University of Pennsylvania, and Skeyes Unlimited. The overall goal of the RCTA was to enable high speed autonomous ground mobility in unstructured environments. Secondary goals included increasing the level of technical expertise within ARL, transitioning research results into Army applications, and contributing to the Army's Future Combat System (FCS) program.

*Arturo.Rankin@jpl.nasa.gov; phone (818) 354-9269; fax (818) 393-5007; <http://robotics.jpl.nasa.gov>

Three technology areas critical to the development of future autonomous unmanned systems were addressed under RCTA: advanced perception, intelligent control architectures and tactical behaviors, and human-robot interaction. The Jet Propulsion Laboratory (JPL) participated as an RCTA member for the entire program, working four tasks in the advanced perception technology area: *Improvements to Stereo*, *Terrain Classification*, and *Pedestrian Detection in Dynamic Environments*, and *Long Range Vision for Terrain Classification*.

The stereo task ran from June 2001 through September 2006. During the first year, we developed a multi-baseline, portable, 5-camera stereo system to compute reference ground truth range estimates [1]. In the remaining years, we focused primarily on algorithmic improvements to each stage of the stereo pipeline. Near the middle of the stereo task, an obstacle detection performance evaluation was performed on a surveyed obstacle course using an early version of RCTA stereo that included the stereo correlator enhancement described in the next section [2]. At the end of the stereo task, terrain maps generated with full and multi-resolution stereo range data were qualitatively evaluated [3]. Stereo vision was used as the front-end for all of JPL's other RCTA tasks. Section two of this paper summarizes JPL's stereo vision contributions to RCTA.

Our terrain classification task ran the full duration of the program. During the first 15 months, we focused on discriminating soil and rock terrain types from vegetation using thermal infrared imagery. Terrain classification with a single spectral band thermal infrared sensor was not robust in all circumstances. However, tests with two spectral bands, mid-wave infrared (MWIR) and long-wave infrared (LWIR), indicated they have potential to discriminate soil/rock from vegetation without separating temperature from emissivity [4]. We expect increased effectiveness with additional spectral bands, however, sensor cost would likely be problematic for most UGV programs. From 2003 until the end of the program, the primary focus of our terrain classification task was water detection. In 2006, mud detection was added to the scope of the terrain classification task and investigated at a slightly lower priority. Section three of this paper summarizes JPL's water and mud detection contributions to RCTA.

In October 2007, a new terrain classification task was initiated to address the issue of classifying terrain at long range using stereo vision and proprioceptive sensors. Under this task, we built on the learning framework initiated during the DARPA Learning Applied to Ground Robots (LAGR) program to classify terrain at long range that has similar features to terrain already traversed by a UGV. This terrain classification task ran from October 2007 to the end of the program. Section four of this paper summarizes JPL's long range terrain classification contributions to RCTA.

During 2005, there was increased awareness from parallel UGV programs concerning the need for UGVs to operate safely around pedestrians during autonomous navigation. In response to this need, RCTA initiated a pedestrian detection task that involved several RCTA members (GDRS, JPL, CMU, UMD, Sarnoff) and a variety of sensor technologies (e.g., laser, monocular vision, stereo vision, structure from motion, thermal infrared). The JPL pedestrian detection task ran from October 2005 to the end of the program and focused on using color stereo cameras. To improve the detectability of partially occluded pedestrians and reduce pedestrian false alarms, a vehicle detection algorithm was developed under our pedestrian detection task. Section five and six of this paper summarizes JPL's pedestrian and vehicle detection contributions to RCTA, respectively.

2. STEREO IMPROVEMENTS

Stereoscopic vision remains the most widely applicable option for passive 3D sensing where a minimal detectable signature is required. Our stereo task focused primarily on improving the accuracy of range estimates generated using stereo images of complex outdoor scenes. Complex scenes have repetitive patterns that lead to mismatches or weak correlations on thin objects such as tall grass, leaves, and branches. Mismatches and weak correlations both result in reduced stereo data density. As a result, small but hazardous obstacles (such as thin trees) that may be detected only at close range. During RCTA, stereo improvements came from use of multiple baselines, higher resolution images, multi-resolution processing, and from algorithmic improvements to each stage of the stereo pipeline. Other activities performed under this task included developing a method to compute reference ground truth range data for stereo performance evaluations, and evaluating obstacle detection algorithms using the stereo enhancements implemented under this task. In this section, we focus on the improvements made to our stereo algorithm.

The algorithmic improvements implemented during RCTA include, and are primarily reflected in, the quality of the intermediate data products. These qualities propagate along the pipeline to produce a higher quality final product, the stereo range data. Starting with the use of high quality lenses, the increased sharpness and contrast of input imagery was the basis for obtaining higher quality output from each step of the stereo pipeline:

1. Stereo rectification: Input images from the left and right cameras are resampled when they undergo a warping transformation to produce rectified images. The resampled images can suffer aliasing and smoothing effects with bilinear methods. Bicubic resampling, however, produces resampled images that have reduced aliasing and has a sharpening effect. This enhancement has resulted in better correlation quality and increased range data density.
2. Photometric correction: Band-pass filtering that reduces the image pair intensity differences and enhances underlying image texture for stereo matching can be accomplished by LoG (Laplacian-of-Gaussian) convolution (for less complex, open terrain scenes) or by more general background-subtraction methods. Our enhanced stereo algorithm uses a bilinear filter that does not smooth across image intensity edges; it combines domain and range filtering thereby enforcing geometric and photometric locality [5][6]. The net effect is a significant decrease in the ringing artifacts that typically occur with box filtering approaches. Using this edge-preserving technique to reduce the aliasing effects of ringing has resulted in better definition of objects in a scene and increased data density.
3. Stereo pyramid construction: The ability to generate stereo range data at rates that satisfy real-time constraints and the flexibility to allow using much higher resolution images is accomplished by processing the stereo pairs at a desired level or levels of an image pyramid. For expediency, block averaging methods have been used to decimate the images to lower resolution levels. To improve the quality of the image pyramid we used Gaussian sampling [7]; it greatly reduces blockiness and aliasing, both contributors of spurious range data.
4. Overlapped correlation: Matching images acquired from different viewpoints leaves occluded portions where range discontinuities occur. The improvements in the previous steps of the pipeline increase data fidelity to begin with, but the fact that a correlation window is necessary for matching produces uncertainties along depth discontinuities. The use of overlapping windows [8] that are simultaneously collecting matching information on both sides of a depth discontinuity significantly reduces the ambiguity between foreground and background objects. In particular, the use of five overlapping windows enabled better assessment of the correlations by also examining the correlations to the left and right, and above and below the current location undergoing matching.
5. Disparity data filtering: Repetitive patterns in the scene, low texture regions, small non-stationary features (e.g. tree leaves), and propagated aliasing generate small regions of spurious range data that must be removed by filtering operations. Besides traditional curvature filters, a blob filter that examines the gradient of the disparity data gives an indication of spurious matches. The improvement here was to measure relative gradients instead of absolute ones; larger gradients are allowed in the nearfield data than in the farfield data.
6. Range data filtering: Remaining foreground/background ambiguities manifest themselves as “mixed” pixels. Small sets of pixels along object borders that are not committed to belong to the object surface or to the background surface. This filtering is accomplished by explicitly looking for these pixels along the light rays [9]. For a given ray only one pixel is allowed in 3D. More than one range pixel along a ray signifies a string of mixed pixels. The filter simply examines trios of range pixels and tests for co-linearity with the camera rays.

Fig. 1 to Fig. 4 illustrate the result of these algorithmic improvements applied to complex scenes using selected combinations of image resolution, camera baseline and correlation window size. The scene in Fig. 1 from the Angeles National Forest contains trees, underbrush, a creek, and a hill in the background. The elements of our pre-RCTA stereo pipeline was bilinear resampling for rectification, a box filter for photometric correction and texture enhancement, 2x2 block averaging for image decimation, single-window sum-of-absolute-differences (SAD) correlation [10], and absolute gradient measurements for blob filtering of spurious range data. Fig. 1 contains the rectified left image of the scene, a pre-RCTA stereo range image, and a stereo range image after improving the correlator. The closest tree is 5.6 meters away and the hill slope in the background is 45.2 meters away. The input images were processed at a resolution of 400x300 pixels, the stereo baseline was 30cm, and the correlation window was 7x7 pixels. In the improved correlator, five overlapping windows (SAD5) are used where the matching score for a pixel is given by the SAD score of a center window plus the best two SAD scores of four surrounding correlation windows.

Fig. 2 shows improved results using higher resolution images. The 4-fold increase in image dimensions (16-fold in area) suggested a 2-fold increase in correlation window dimensions to increase data density while avoiding smoothing effects. A comparable factor in window size increase (from 7 pixels to 27 pixels) would increase data density further at the expense of the added accuracy in the subpixel range estimates gained by the smaller correlation window sizes. A 16-fold increase in image area significantly increases the computational cost of the stereo correlator, especially using SAD5. Significant effort was spent in optimizing the correlation step, the highest cost component of the stereo pipeline, to maintain real-time performance. JPL produced FPGA implementations of standard SAD and overlapping SAD5

correlators under a parallel UGV programs. However, processing even higher resolution images required applying a multi-resolution approach, described later in this section.

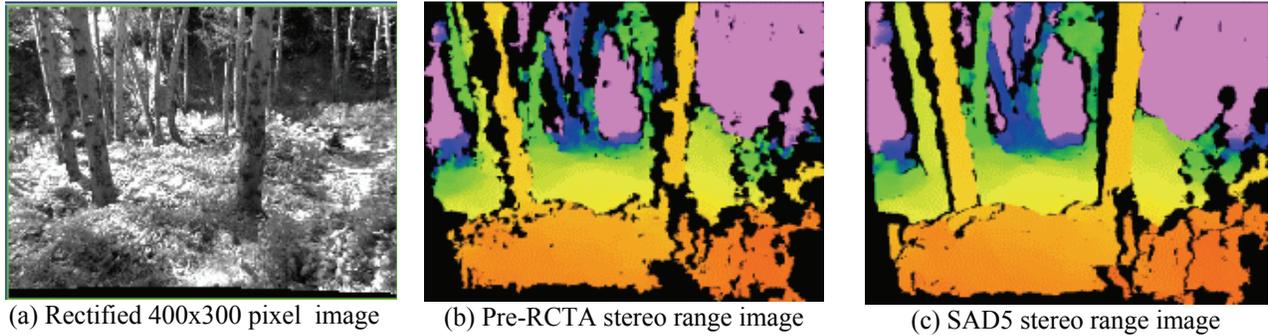


Fig. 1. Improved stereo range data quality at low resolution. The stereo baseline is 30cm and the correlation window is 7x7 pixels. The range data in (b) is from our pre-RCTA stereo algorithm. The range data in (c) was generated after implementing the SAD5 correlator. Red is closer, magenta is far. Note the increased fidelity of the tree shapes.

Fig. 3 illustrates improved results at higher resolution but using a reduced stereo baseline. The correlation costs are reduced from matching a narrower range of disparities but the range accuracy is reduced. Small baselines produce more detail and less matching errors for smaller objects like leafy bushes and thin branches. Fig. 3a shows the loss in range accuracy downrange for 30cm and 6cm baselines for the range data results in Fig. 3b and Fig. 3c. The difference between these two results is the application of the mixed-pixel filter. Note that the edges of the trees and bush leaves appear black (no data), indicating a clearer separation between foreground and background surfaces.

Processing even higher resolution images in real-time suggests processing swaths of the image at different resolutions. The increased detail in the near field affords processing at lower resolutions while the far field requires the increased data density and fidelity. One approach is to determine the boundaries of the swaths so that a maximum horizontal pixel size is maintained across the resolution pyramid. Fig. 4 shows a 1600x1200 image of a scene containing a dirt road with a 30cm wide, 20cm tall rock on the road surface about 16m away. Maintaining a 2cm horizontal pixel size, the rock falls on the half-resolution (level 1) swath. (Full resolution is level 0, 1600x1200 images, with each level halving the resolution). Several observations can be made regarding the level 3 swaths. Firstly, the level 3 swaths cover almost 2/3 of the images. Secondly, processing the level 3 swaths produces adequate range data for planning purposes over the near field and the region above the road surface. Finally, processing at level 3 where possible has low cost and would provide information on the disparity search ranges appropriate for the type of scene, driving speed, and look-ahead monitoring.

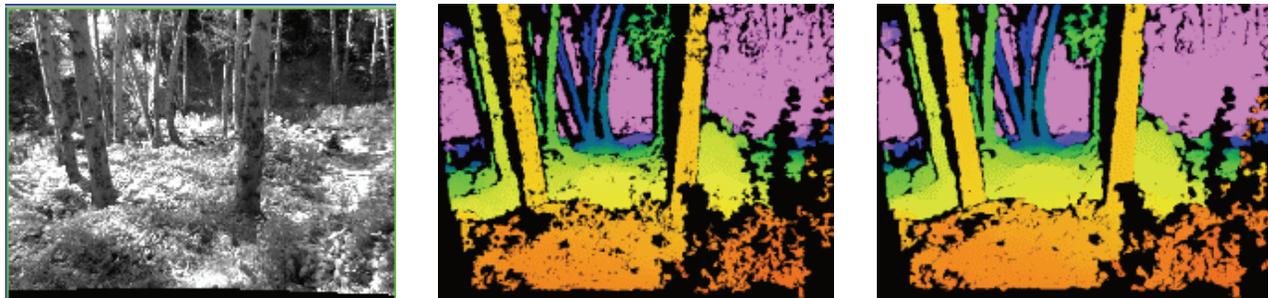
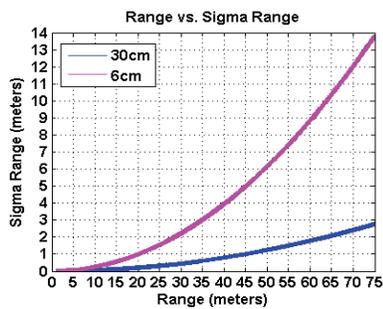
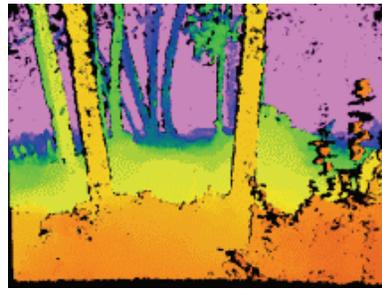


Fig. 2. Improved stereo range data at higher resolution. The stereo baseline is 30cm and the correlation window is 13x13 pixels. The range data in (b) and (c) has undergone mixed-pixel filtering. Note the green tree on the left is separated from the occluding, closer tree. Note also the increased range data density in (c).

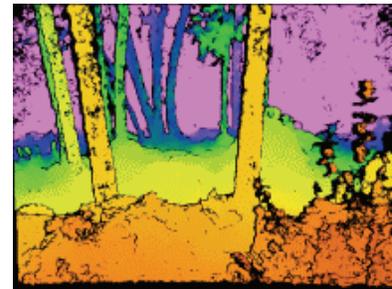
Fig. 2. Improved stereo range data at higher resolution. The stereo baseline is 30cm and the correlation window is 13x13 pixels. The range data in (b) and (c) has undergone mixed-pixel filtering. Note the green tree on the left is separated from the occluding, closer tree. Note also the increased range data density in (c).



(a) Range data accuracy for 6cm and 30cm stereo baselines. A function of the square of the range, the match precision, the field of view of the cameras and the stereo baseline.

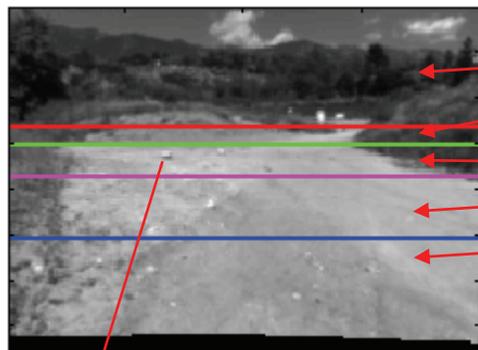


(b) Range image after all improvements have been applied but prior to mixed-pixel filtering. Such filters can be turned-off depending on the scene characteristics.



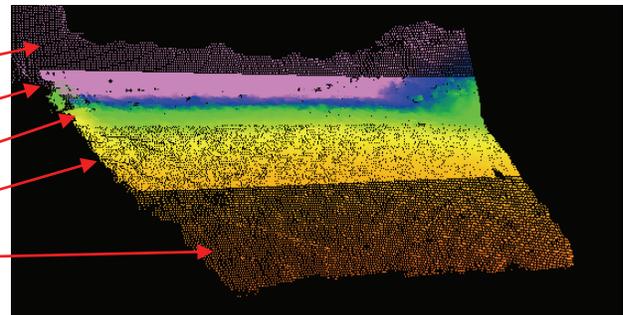
(c) Range image after all improvements have been applied. The 6cm baseline reduces the cost of correlation but has limited range accuracy at the longer ranges.

Fig. 3. Improved range data quality at higher resolution. The stereo baseline is 6cm and the correlation window is 13x13 pixels. The range data in (c) has undergone mixed-pixel filtering. Note the “green” tree on the left is separated from the occluding, closer tree in (c).

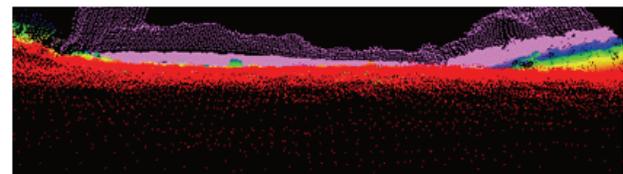


a) Image swaths illustrated on the level 3 image. Each swath maintains a constant 2cm pixel size in this example. Swath boundaries can be chosen to match a scene type, driving speed and need to anticipate moderate to far-field presence of hazards. The rock illustrated is 30cm x 20cm tall, and is located about 16 meters away.

Level 3
Level 0
Level 1
Level 2
Level 3



b) 3D view of distance-coded range data from above and left. Orange is closer. The range data is combined from the swaths and is seamless, with enough density for near, intermediate and far-field interpretation and reactions.



c) 3D view of range data from surface level. Height-colored range data. Red is lower, ground level. Magenta is higher terrain. The rock is well detected at level 1.

Fig. 4. A multi-resolution scheme can process the high resolution images at reduced cost. The image pyramid consists of 4 levels. Terrain hazards can be detected at a distance to afford reaction times. The cost of processing for the parameter chosen (2cm constant pixel size in each swath) is about 0.8x the cost of processing the level 2 (400x300) image in full, 8x for the full level 1 image (800x600) and 64x for the level 0 (1600x1200) image.

3. TERRAIN CLASSIFICATION: WATER AND MUD

Robust water detection is a critical perception requirement for cross country unmanned ground vehicle (UGV) autonomous navigation. Traversing through deep water bodies could cause costly damage to the electronics of UGVs. Additionally, a UGV that is either broken down due to water damage or stuck in a water body during an autonomous military mission may require rescue, potentially drawing critical resources away from the primary mission and soldiers into harm's way. In this section, we summarize our approach to detecting water hazards during the daytime (using a stereo pair of color cameras) and localizing them in world model from which a UGV can perform autonomous navigation.

The appearance of water can greatly vary, depending upon the color of the sky, the level of turbidity, the time of day, and the presence of wind, terrain reflections, underwater objects visible from the surface, surface vegetation, and shadows. The large number of possible scenarios and appearances of water makes its detection particularly challenging using a single cue. Early in the program we developed a multi-cue water detector that uses a rule base to combine water cues from color, texture, and stereo range reflections [11]. Hue, saturation, and brightness levels are thresholded to generate the water cue from color. These thresholds are tuned to detect sky reflections in water. Local image intensity variance is thresholded to generate the water cue from texture. Stereo data is analyzed to detect patterns consistent with reflections of objects (such as trees). There are two complementary water cues from stereo data. First, non-zero stereo disparity on reflections of objects in still water will have a range nearly identical to the range of the reflecting object. Second, zero stereo disparity that occurs below the horizon may be evidence of reflections of object in still water that are extremely far away, such as clouds or distant tree lines. (Zero stereo disparity occurs when the stereo correlator matches the same column in rectified left and right images.) Fig. 5 contains a sample multi-cue water detection result.

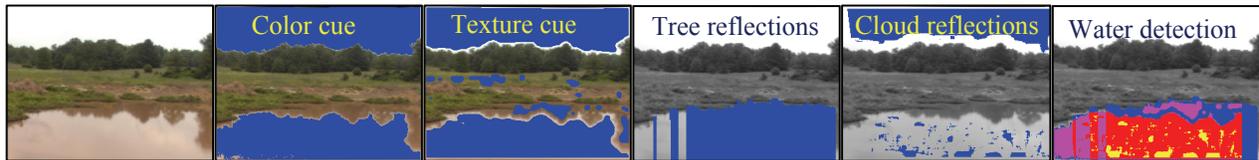


Fig. 5. A multi-cue water detection approach was implemented early in the RCTA program. A rule-base was used to merge cues from color, texture, terrain reflections, and cloud reflections. In the fused water detection image, water cues above the horizon are ignored. (blue = 1 cue, magenta = 2 cues, red = 3 cues, yellow = 4 cues).

In the later half of the program, we developed three stand alone water detectors, each exploiting different characteristics of water. In practice, we run all three water detectors in parallel and union the output of each detector. The first detector was designed to locate water bodies in cluttered environments by detecting reflections of object in stereo range data. Reflections of ground cover (such as trees and vegetation) extend from the trailing edge of a water body and can span a portion or all of the water body, depending upon the reflected body's height and distance from the water. In stereo range data, the range to a reflection roughly matches the range to the reflecting object, however, the elevation of a reflection is significantly lower than that of the surrounding ground plane. As these are distinct features, object reflections are straightforward to segment in stereo range data. The elevation of the perimeter of a detected object reflection is averaged and used to correct the object reflection range data so that it corresponds to the surface of the water body. Water detection results and corrected stereo range data are used to generate a 2.5D world map [12]. Fig. 6 illustrates stereo reflection based water detection and localization in the world map.

The second detector was designed to locate water bodies that are out in the open on cross-country terrain at mid- to far-range (approximately 20-60 meters). Non-traversable water bodies, such as large puddles, ponds, and lakes, are indirectly detected by detecting reflections of the sky below the horizon in color imagery. The appearance of water bodies in color imagery largely depends on the ratio of light reflected off the water surface to the light coming out of the water body. When a water body is far away, the angle of incidence is large, and the light reflected off the water surface dominates. We have exploited this behavior to detect water bodies out in the open at mid- to far-range. When a water body is detected at far range, a UGV's path planner can begin to look for alternate routes to the goal position sooner, rather than later. As a result, detecting water hazards at far range generally reduces the time required to reach a goal position during autonomous navigation. This detector geometrically locates the exact pixel in the sky that is reflecting on a candidate water pixel on the ground, and predicts if the ground pixel is water based on color similarity and local terrain features.

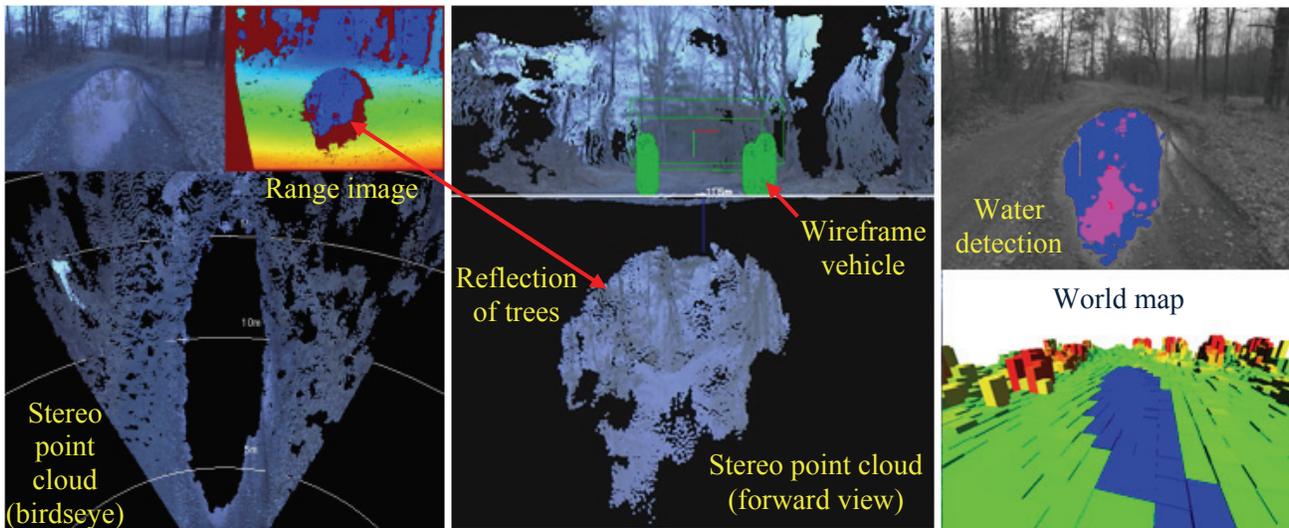


Fig. 6. In cluttered environments, water can be indirectly detected by detecting reflections of objects. Stereo range data corresponding to reflections of objects in water bodies will lie under the ground surface and have a range close to the range of the actual object being reflected. In the water detection image, reflections of trees (blue) and the sky (magenta) were detected. In the world map, blue represents detected water, green represents highly traversable terrain, red represents non-traversable terrain, and yellow indicates an intermediate level of traversability.

When water bodies are out in the open and at far range, sky reflections provide a strong cue for water. But at close range, the color coming out of a water body dominates sky reflections, and the water cue from sky reflections is of marginal use. At close range, the color of a water body tends to gradually change from the leading edge to the trailing edge. Brightness and saturation levels tend to increase, with saturation content changing at a faster rate than the brightness content. For all the pixels on a water body, a plot of brightness/saturation vs. incidence angle is fairly linear with high slope. Fortunately, this slope tends to be higher for water than other naturally occurring terrain. We have exploited this phenomenology to develop a third detector that locates water bodies at close range based on the variation in color across the water body. First, we identify candidate water regions in image space by locating regions having low texture. Next, we evaluate the color changes across each candidate water region to locate those consistent with water. Finally, we perform an ellipse fit on remaining candidate water regions and apply some size and aspect ratio filtering to prune regions that geometrically are not likely to be water. The sky reflection and color variation based water detectors are complementary. The color variation based water detector performs well in detecting water bodies at close to mid range, and the sky reflection based detector performs well in detecting water bodies at mid to far range. Fig. 7 illustrates combined sky reflection and color variation based water detection on a ground-truthed sequence containing 143 frames.

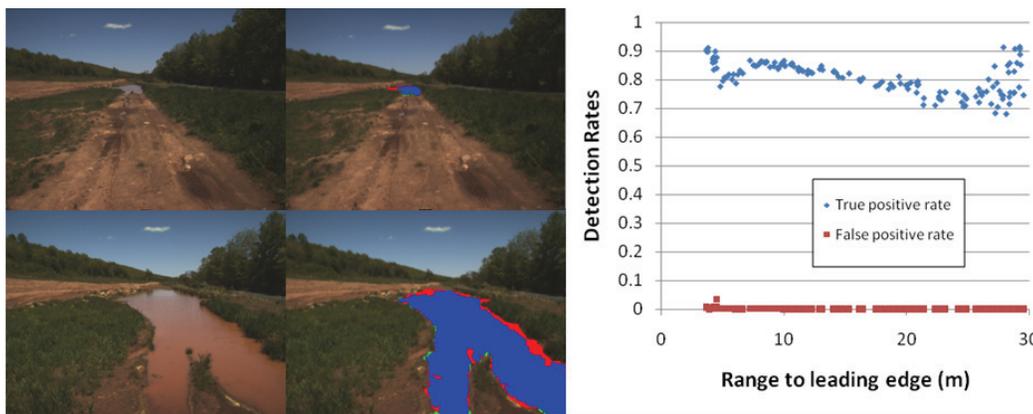


Fig. 7. Water detection results for the first and last image in a ground-truthed sequence containing 143 frames. The blue, red, and green pixels indicate true positive detection, false negative detection, and false positive detection, respectively.

The JPL library of water detectors was integrated with the GDRS autonomous mobility (AM) [13] and C2 global map and planners (Fig. 8) and evaluated by ARL during the program ending capstone experiment. Water detection (including stereo ranging at a resolution of 512x384) runs at 1Hz on a 3GHz Pentium Core 2 Duo processor.

During our mud detection effort, JPL performed daytime data collections on wet and dry soil using a variety of passive imaging sensors that span much of the visible thru thermal infrared portion of the electromagnetic spectrum. These include color stereo, multi-spectral (visible plus near-infrared), short-wave infrared (SWIR), MWIR, LWIR, and polarization. In [14], we characterized the strengths and weaknesses of each passive sensor type for detecting mud under nominal conditions. Our analysis indicated that polarization cameras are particularly useful for segmenting mud surrounded by dry soil since the degree of linear polarization (DOLP) on mud is consistently higher than surrounding dry soil, regardless of sky conditions, the sun position, and the sensor orientation [14]. A mud detection algorithm was implemented which thresholds DOLP and back-projects polarization pixels that have a high DOLP into the corresponding rectified left color image (which is registered with the stereo range image). As we only expect mud to exist on the load bearing surface, stereo range data is used to reject false detections on the parts of a scene that is not a part of the ground. A sample mud detection result is illustrated in Fig. 9.

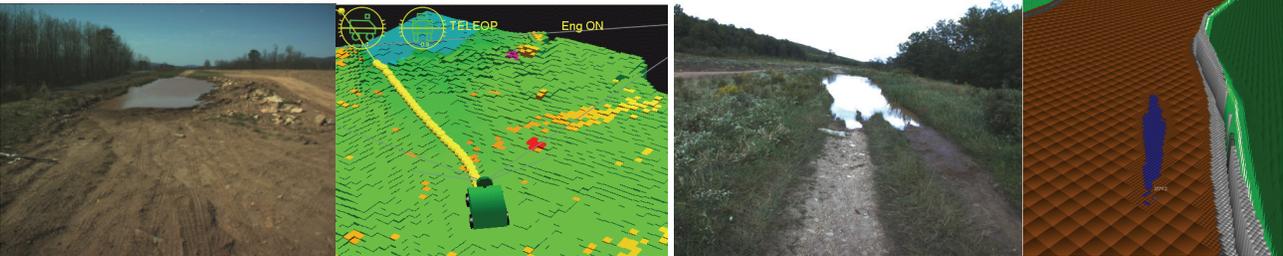


Fig. 8. The water detector was integrated with the GDRS AM map and planner (left) and the GDRS C2 global map and planner (right). The AM map has a resolution of 40cm and is vehicle centered, extending 20 meters in each cardinal direction. The C2 global map, which is used to plan up to 3-5km, has a resolution of 1 meter where water has been detected and 5 meters where no sensed data exists. The blue regions in each map represent detected water. The yellow planned path in the AM map is avoiding the detected water.



Fig. 9. Sample mud detection using a stereo pair of color cameras and a SALSA polarization camera. Polarization pixels with a high DOLP are back projected into the left rectified color image, which is registered with stereo range data. Since mud only occurs on the load bearing surface, stereo range data is used to reject false detections on ground clutter.

4. TERRAIN CLASSIFICATION: UNDERFOOT TO FAR FIELD LEARNING

Traditional off-road autonomous navigation algorithms for unmanned ground vehicles have used fixed terrain traversability evaluation based on programmed rules or classifiers learned with operator supervision. However, this results in brittle behavior when novel terrain is encountered. To be truly robust, autonomous vehicles must be able to adapt to the terrain and learn from experience. Under the RCTA program, we have continued to complete the end-to-end learning framework proposed in [15] under the LAGR program. We started by transferring the ability to learn long-range visual terrain classification from binary stereo-based classification on the LAGR vehicle to an experimental unmanned vehicle (XUV) driving at speeds up to 30 kph. We then addressed the problem of learning the mid-range classification from vehicle proprioceptive sensors, with no operator supervision. We also extended the terrain classification from binary classes (traversable and non-traversable) to multiple, automatically learned classes with different degrees of traversability.

4.1 Long Range Binary Terrain Classification

In order to classify terrain at long ranges, where range data is too sparse to apply standard geometric terrain classification, we learn the appearance of terrain classes in the near and mid fields based on the geometric classification. This allows us to classify the terrain at significantly longer ranges than a standard geometric classifier. Our long-range terrain classification algorithm consists of a training and classification stage. The training stage consists of computing dense stereo, projecting the stereo data into a polar-perspective map to classify terrain based on geometry, back-projecting the vehicle's path into the image to determine regions that can be associated with proprioceptive data, extracting color histogram features for the image, and then training a linear support vector machine (SVM) classifier based on the examples determined from geometry and proprioception. Depending on the number of features being used for training, this process can take 0.5-1.0 seconds for 512x384 resolution imagery on the RCTA's XUV onboard computer. The classification stage only requires extracting features from a single image in the regions where terrain classification will be performed (generally only the top half of the image for long range terrain classification) and the classification of these features by the SVM. Because the SVM results in a linear decision surface, the classification of features is extremely fast, and the computation time is dominated by feature extraction. On the XUV computer, computing the color histogram features for each pixel in the top half of a 512x384 image takes approximately 0.1 seconds. The training process periodically updates the classifier, which is then used for classification of any new images acquired until a new update is provided.

To analyze the performance of the classification, we compare the classification error between the system's long-range appearance-based classification and the actual path taken by the vehicle and the classification from geometry. While this only provides the classification error in specific regions of the image, it generally provides a good indication of the overall performance of the system and specific failure modes (such as when the vehicle drives into terrain with a very different appearance or turn sharply into or out of the sun). However, in general, after driving over the complex terrain (mixed concrete and grass surface, with cars, pedestrians, bushes, and trees), the system is able to classify most of the frames from the sequence correctly, as shown in Fig. 10.

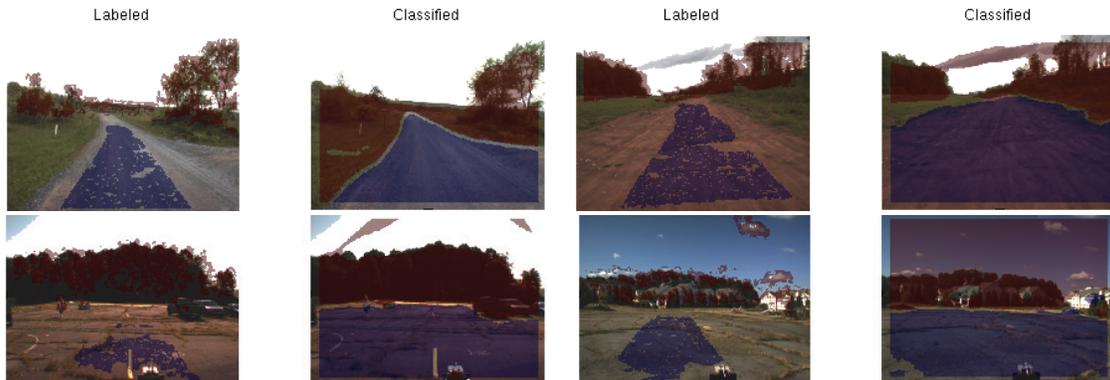


Fig. 10. Examples of terrain classified at long-range by learning the appearance of terrain from the vehicle path and stereo geometry; the blue labels represents traversable terrain, while the red labels represent non-traversable terrain.

4.2 Unsupervised, Multi-Class Terrain Classification

The long-range terrain classification system originally associated the visual appearance of terrain that the vehicle drove over with the "traversable" class. However, it did not take into account any proprioceptive data when driving over the terrain, and therefore could not distinguish the degree of traversability of the terrain. Unfortunately, terrain traversability is an extremely complex measure, depending on the vehicle and terrain properties and the vehicle-terrain interaction. As such, rather than attempting to model traversability directly, we simply attempt to find clusters within the proprioceptive data. These clusters then represent the expected response of the vehicle while driving over a certain class of terrain. After the clusters have been determined, the visual appearance of the clusters can be used to learn a discriminative classifier, which can then be used to predict the vehicle response from visual input.

Previous work has shown that the vibration measured at the wheels and accelerations from an inertial measurement unit (IMU) can be used to distinguish different terrain classes. However, these approaches relied on supervised training in controlled settings to train a classifier. We have extended these approaches to automatically learn terrain classes during any type of driving (autonomous or tele-operated). To exploit the spatial coherency of terrain types as well as the vehicle response, we use a hidden Markov models (HMMs) and spectrogram features of IMU pitch rates to separate the time series of pitch rate data into segments with similar behavior characteristics.

Specifically, we use selected frequency components of a sliding-window fast Fourier transform (FFT) of the incoming sensor stream $x(t)$ to derive a time series of vectors $y(t)$. These vectors explicitly encode slow, rolling motions versus fast, choppy motions. This stream of vectors is separated into a corresponding sequence of discrete classes $z(t)$ by selecting a HMM using an optimization technique which maximizes goodness-of-fit subject to constraints.

After learning the terrain classes, dense stereo information is used to back-project the vehicle's classified path into imagery, from which a multiclass SVM visual classifier is learned. This classifier is a standard one-against-all extension to the one used in the binary case, with the same color histogram features. Because the system uses unsupervised learning of terrain classes, the vehicle can learn continuously throughout its life. In particular, the vehicle can exploit large amounts of logged data to learn visual classes that are invariant to illumination and environmental conditions, because they include all observed variations. Sample terrain classification results are shown in Fig. 11.



Fig. 11. Top row: Examples of terrain classified by unsupervised classification using proprioceptive sensors, where different color labels represent different terrain classes, and the striped labels represent areas of low confidence. Bottom row: Example of terrain classified by using the classes learned by unsupervised proprioceptive learning as self-supervision to an appearance-based classifier.

5. PEDESTRIAN DETECTION

Detection of personnel or pedestrians in the vicinity of an unmanned ground vehicle is motivated by the requirement to operate safely and effectively while navigating a cluttered urban environment or primary or secondary roads. Typical driving speeds in these situations dictate a detection range of 30-60m, with a false alarm rate of only several per hour. Detection at even longer ranges is useful for surveillance from a moving vehicle. Furthermore, effectively planning trajectories among people requires them to be localized and that their motion be predicted by tracking them over time. Detecting people is challenging due to their variety of pose and appearance, along with clutter and occlusion in the scene. Motion of a person cannot be relied upon solely, as stationary pedestrians are equally in danger to those walking.

We address the problem of reliably detecting upright people from a moving vehicle in cluttered, dynamic scenes out to a range of 40m with a stereo vision based approach. Stereo vision has the advantages of providing passive high resolution imagery co-registered with dense range data, in a low-power, low-cost package. We utilize shape information from dense stereo processing to find regions of interest, eliminating the need for a sliding window search, and remove the background in the regions. We then compute both shape and appearance features for the foreground pixels of the regions and use them to classify the region with a probability of being a human. Because the range to the target is known, our approach does not require a search over multiple scales, greatly reducing its computational requirements.

Our approach [16][17][18] adopts a processing pipeline similar to many other pedestrian detection systems, but with the key difference that we exploit dense stereo at all stages to improve classification and runtime performance. After computing dense stereo, we perform visual odometry to compute the change in pose of the vehicle, detect regions of interest in the current frame, classify each region with a probability of being human, and then track the regions over multiple frames, declaring the track a human after filtering the classification.

To detect regions-of-interest and segment potential pedestrians from the background, we project the stereo range data into a two-dimensional polar perspective grid map (PPM), whose cells are defined with a fixed angular resolution, but variable range resolution to preserve the coherency of the stereo range data. The number of range points is accumulated into cells and the map is smoothed with an averaging filter. The map gradient is then used to segment clusters of points, which are then merged into objects based on a fixed size threshold.

Geometric and visual features of each segmented 3D point cloud are used to classify them as human or not human based on shape. The geometric features include the fixed-frame shape moments (variances of point clouds in a fixed frame), rotationally invariant shape moments (the eigenvalues of the point cloud's scatter matrix), and "soft-counts" of various width, height, depth, and volume constraints. To compute the visual features, the segmented region of interest is scaled to a canonical size image mask. Visual features are then computed as the moments of a binary, edge, and color foreground mask. Logarithmic and empirical logit transforms of the moments and counts are used to improve the normality of the feature distribution. We then expand the features set to use higher-order terms in a linear classifier to maintain computational efficiency and robustness. Using this feature vector, a Bayesian generalized linear model (GLM) classifier (for logistic regression) is then trained using standard iteratively reweighted least squares (IRLS) to obtain a Gaussian approximation to the posterior mode. Simple MAP estimates of predictive probability (of being human) are obtained using this Gaussian mode-based approximation.

Tracking is used to estimate the velocity of objects while also reducing incorrect detections by filtering classification probabilities. Regions-of-interest are matched between consecutive frames by first gating neighboring regions based on distance, and then selecting the minimum cost based on the regions' color histograms. The classification probabilities are filtered with a median of three consecutive scores and the velocity of objects is estimated by fitting a linear motion model to the track.

The system has been tested in highly cluttered, urban environments and on less cluttered, secondary roads. Results on approximately 4400 frames of data with 3400 ground-truthed pedestrians on lightly cluttered secondary roads are presented in Fig. 12. Examples of performance in highly cluttered, urban environments are shown in Fig. 13. The entire system runs at approximately 7Hz on 1024x768 imagery, with the detection, classification, and tracking portions taking 100ms total on a standard 2.4GHz Intel Core 2 Quad processor.

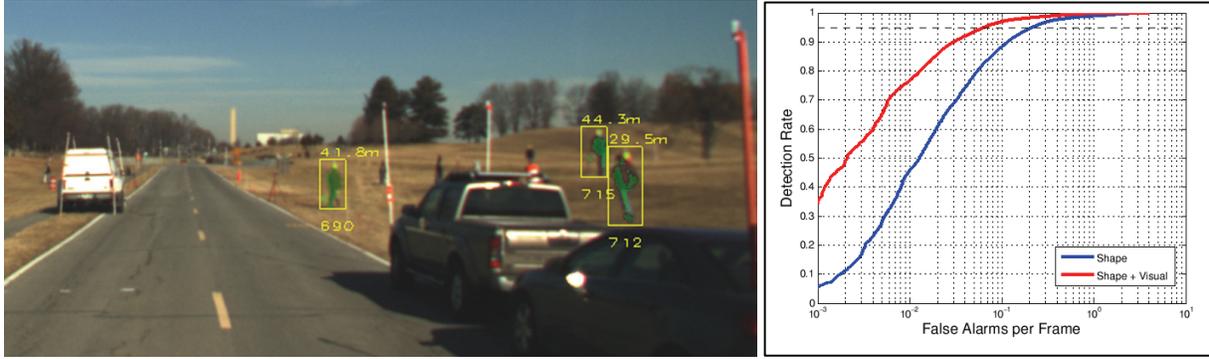


Fig. 12. Performance on lightly cluttered secondary roads; the detection rates are shown for 1000 trials of 80/20 split cross-validation tests of data from ~4400 frames with ~3400 annotated pedestrians for the case of using only geometry features, and using both geometry and appearance features.



Fig. 13. Examples of detection performance on highly cluttered urban scenes. Yellow bounding boxes show correct detections, with a green overlay of the pixels belonging to the pedestrian; cyan boxes show missed detections.

6. VEHICLE DETECTION AND CLASSIFICATION

Pedestrians who are partially occluded are particularly difficult to detect. We have observed that in urban areas, partially occluded pedestrians (and pedestrian false detections) largely occur due to vehicles. Therefore, explicitly detecting vehicles has a high utility as a pre-processing stage to estimate where occlusion may occur and where people should not be found. JPL developed a two stage algorithm for finding vehicles in stereo imagery. In the first stage large blobs in the stereo point cloud are found by projecting the data to a local-ground-plane map and looking for peaks in the map. Detected blobs are then assigned a feature vector which describes the blobs shape. Finally, the blobs feature vector is passed to a linear classifier which determines whether or not the blob is a vehicle. Each detected vehicle has an associated binary mask which denotes which pixels in the image correspond to that vehicle. Each detected vehicle is also assigned a 3D coordinate in the cameras coordinate frame.

Our vehicle detection and classification algorithm has been tested using RCTA stereo imagery collected at GDRS (Westminster, MD) and the National Institute of Standards and Technology (Gaithersburg, MD). A sample result is shown in Fig. 14. When RGB images having a resolution of 1024x768 pixels are processed, the algorithm has a processing time of 16 milliseconds per frame (not including stereo computation) on an Intel Core2 Quad CPU clocked at 2.40 GHz, and consumes 44 megabytes of memory. The algorithm is able to detect vehicles at ranges up to 60 meters. ROC curves showing the probability of detection as a function of the number of false alarms per frame (FAPF) are given in the Fig. 15. There are multiple curves in order to show the performance of the system at different ranges.

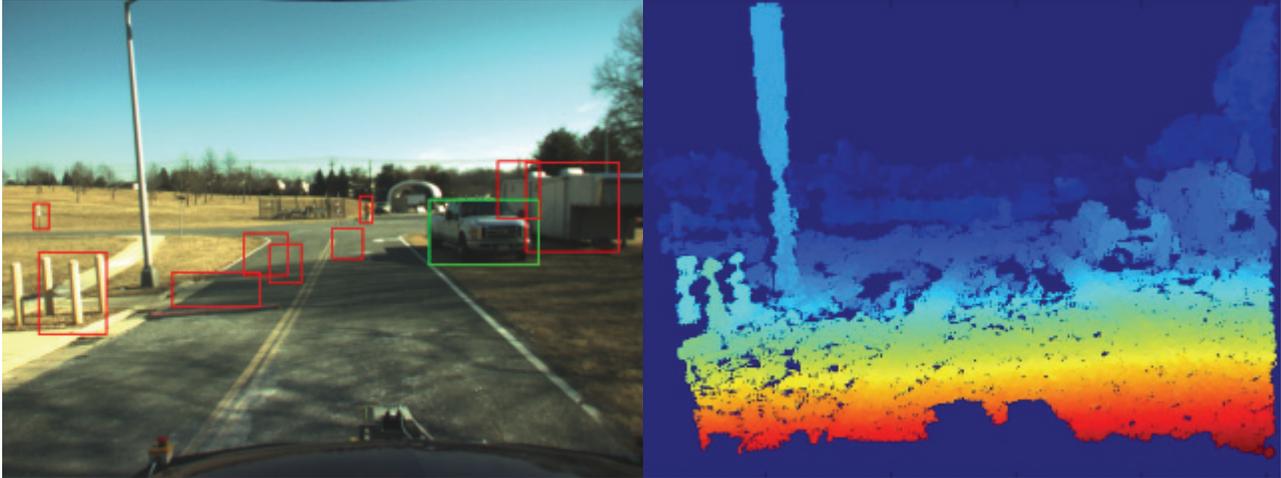


Fig. 14. Sample vehicle detection from a moving vehicle for a cluttered scene (left) using stereo range data (right). The boxes in the scene identify candidate vehicle blobs that were extracted from the stereo range data. The green box contains a blob correctly classified as a vehicle and the red boxes contain blobs correctly classified as something other than a vehicle. In the stereo range image, the red pixels correspond to close range, blue pixels correspond to far away, and the colors between red and blue correspond to an intermediate range.

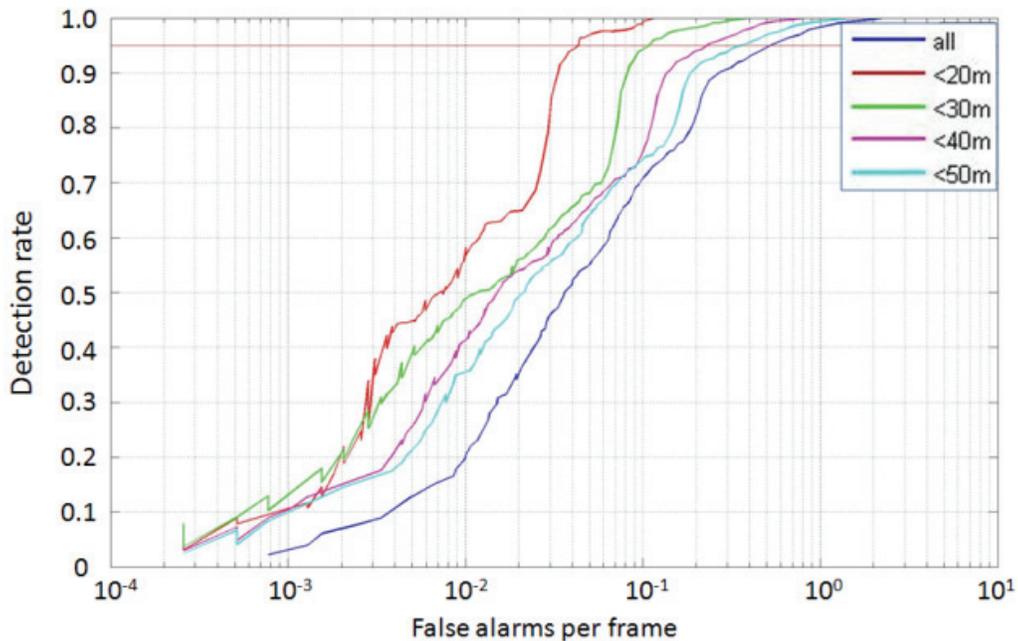


Fig. 15. ROC curves showing the probability of vehicle detection as a function of the number of false alarms per frame for vehicles within 20, 30, 40, and 50 meters of the stereo sensors. Approximately 1200 stereo frames were processed.

7. CONCLUSIONS

JPL participated in the eight year RCTA program, working four tasks in the advanced perception technology area: *Improvements to Stereo*, *Terrain Classification*, and *Pedestrian Detection in Dynamic Environments*, and *Long Range Vision for Terrain Classification*. In this paper, we summarized JPL's contributions to the RCTA program. Under our stereo task, improvements were made to each stage of the stereo pipeline. Bicubic resampling was implemented for

image rectification, a bilinear filter was implemented for photometric correction, Gaussian sampling was implemented for image decimation, five overlapping windows were implemented for correlation, relative gradients were implemented for disparity filtering, and a new stage was added to filter mixed pixels from stereo range data. The modifications to each stage of the stereo pipeline have led to improved intermediate data products and the final end product (stereo range data). Since stereo vision was used as the front end for all of our other RCTA tasks, improvements to the quality of stereo range data led to tangible improvements to all of our tasks.

Under our terrain classification task, water detection and mud detection algorithms were developed. For mud detection, data collections were performed with a variety of passive sensors, the strengths and weaknesses of each sensor was characterized, and a multi-sensor mud detector was developed that fuses cues from color stereo and polarization sensors. Early in the program, a daytime multi-cue water detector was developed that fuses cues from color, texture, and stereo. Subsequently, three stand-alone water detectors were developed that are typically run in parallel. The first detects water bodies out in wide open areas at mid to far range based on sky reflections, the second detects water bodies in wide open areas at close to mid range based on the distinctive variation of color across a water body from the leading to trailing edge, and the third detects water bodies in cluttered environments based on reflections of objects in the background (such as trees). The algorithm is able to detect water bodies at ranges up to 60 meters. During the program ending capstone experiment, JPL's water detection algorithm was integrated onto two XUVs and evaluated at Fort Indiantown Gap (FITG) during autonomous navigation. ARL anticipates publishing results from the FITG experiment later this year [19][20].

Under our long range terrain classification task, we transferred the ability to learn long-range visual terrain classification from binary stereo-based classification on the LAGR vehicle to an XUV, addressed the problem of learning the mid-range classification from vehicle proprioceptive sensors with no operator supervision, and extended the terrain classification from binary classes (traversable and non-traversable) to multiple, automatically learned classes with different degrees of traversability.

Under our pedestrian detection task, we developed an algorithm to reliably detect upright people from a moving vehicle in cluttered, dynamic scenes out to a range of 40 meters using color imagery and stereo range data. Dense stereo is used to select regions of interest based on shape, and to remove the background pixels in the selected regions. Shape and appearance features for the foreground pixels of the selected regions are used to classify them with a probability of being a pedestrian. Selected regions are tracked over multiple frames. During the program ending capstone experiment, JPL's pedestrian detection algorithm was integrated onto a UGV and evaluated at the National Institute of Standards and Technology (NIST), Gaithersburg, MD. ARL anticipates publishing results from the RCTA capstone experiment at NIST later this year [19][20]. ARL plans on publishing results from an earlier experiment at NIST (January 2009) in the proceedings of this conference [21].

To improve the detectability of partially occluded pedestrians and reduce pedestrian false alarms, a two stage vehicle detection algorithm was developed. In the first stage, large blobs in the current stereo range point cloud are extracted and assigned a feature vector. In the second stage, the feature vector of each extracted blob is passed to a linear classifier which determines whether or not the blob is a vehicle. The algorithm is able to detect vehicles at ranges up to 60 meters.

JPL's pedestrian, vehicle, and water detection software has been delivered to the FCS Autonomous Navigation System (ANS) subsystem. JPL's pedestrian detection software has also been delivered to the TARDEC SafeOps program. Under FCS-ANS funding, JPL's SAD5 stereo correlator has been implemented in FPGA and transitioned into FCS-ANS.

ACKNOWLEDGEMENTS

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by ARL under the RCTA program, through agreements with the National Aeronautics and Space Administration (NASA). Reference herein to any specific commercial product, process, or service by trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

REFERENCES

- [1] Ansar, A. and Matthies, L., "Multibaseline stereo system for evaluation of binocular stereo," *Proceedings of SPIE*, Vol. 5083, Orlando, 243-252, (2003).
- [2] Rankin, A., Huertas, A. and Matthies, L., "Evaluation of stereo vision obstacle detection algorithms for off-road autonomous navigation," *Proceedings of the 32nd AUVSI Symposium on Unmanned Systems*, Baltimore, (2005).
- [3] Rankin, A., Huertas, A. and Matthies, L., "Stereo vision based terrain mapping for off-road autonomous navigation," *Proceedings of SPIE*, Vol. 7332, Orlando, 733210-733210-17, (2009).
- [4] Rankin, A., Bergh, C., Goldberg, S., Bellutta, P., Huertas, A. and Matthies, L., "Passive perception system for day/night autonomous off-road navigation," *Proceedings of SPIE*, Vol. 5804, Orlando, 343-358, (2005).
- [5] Ansar, A., Castano, A. and Matthies, L., "Enhanced real-time stereo using bilateral filtering," *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, Thessalonica, Greece, (2004).
- [6] Ansar, A., Huertas, A., Matthies, L. and Goldberg, S., "Enhancement of stereo at range discontinuities," *Proceedings of the SPIE*, Vol. 5422, Orlando, 24-35, (2004).
- [7] Huertas, A., Rankin, A. and Matthies, L., "Improvements to stereo," *RCTA Quarterly Technical Report JPL-03QFY04 AP-04-01-001.doc*, submitted to General Dynamics Robotic Systems, Westminster, MD, July (2004).
- [8] Hirshmueller, H., "Improvements in real-time correlation-based stereo vision," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 141-148, (2001).
- [9] Stein, A., Rankin, A. and Matthies, L., "Improvements to Stereo", *RCTA Quarterly Technical Report JPL-03QFY05-AP-05-01-001.doc*, submitted to General Dynamics Robotic Systems, Westminster, MD, July (2005).
- [10] Matthies, L., "Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation," *International Journal of Computer Vision*, vol. 8, pp. 71-91, (1992).
- [11] Rankin, A., Matthies, L. and Huertas, A., "Daytime water detection by fusing multiple cues for autonomous off-road navigation," *Proceedings of the 24th Army Science Conference*, Orlando, (2004).
- [12] Rankin, A. and Matthies, L., "Daytime water detection and localization for unmanned ground vehicle autonomous navigation," *Proceedings of the 25th Army Science Conference*, Orlando, (2006).
- [13] Lacaze, A., Murphy, K. and DelGiorno, M., "Autonomous mobility for the demo III experimental unmanned vehicles," *Proceedings of the AUVSI Conference*, Orlando, (2002).
- [14] Rankin, A. and Matthies, L., "Daytime mud detection for unmanned ground vehicle autonomous navigation," *Proceedings of the 26th Army Science Conference*, Orlando, (2008).
- [15] Bajracharya, M., Howard, A., Matthies, L., Tang, B. and Turmon, M., "Autonomous Off-Road Navigation with End-to-End Learning for the LAGR Program," *Journal of Field Robotics*, 26(1), (2009).
- [16] Bajracharya, M., Moghaddam, B., Howard, A., Brennan, S. and Matthies, L., "A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle," *International Journal of Robotics Research*, 28(11-12), (2009).
- [17] Bajracharya, M., Moghaddam, B., Howard, A., Brennan, S. and Matthies, L., "Results from a Real-time Stereo-based Pedestrian Detection System on a Moving Vehicle," *International Conference on Robotics and Automation*, (2009).
- [18] Bajracharya, M., Moghaddam, B., Howard, A. and Matthies, L., "Detecting personnel around UGVs using stereo vision," *Proceedings of the SPIE*, Vol. 6962, Orlando, 696202.1-696202.13 (2008).
- [19] Bodt, B., Childers, M. and Camden, R., "A capstone experiment to assess unmanned ground vehicle tactical behaviors developed under the robotics collaborative technology alliance," to appear in *Proceedings of AUVSI*, Denver, (2010).
- [20] Bodt, B. and Childers, M., "A place at the table: the role of formal experimentation in the robotics collaborative technology alliance," to appear in *Proceedings of the 78th MORS Symposium*, Quantico, VA, (2010).
- [21] Bodt, B. and Camden, R., "A final evaluation of pedestrian detection and tracking from a moving platform," to appear in *Proceedings of SPIE*, Vol. 7692, Orlando, (2010).