Detecting Pedestrians with Stereo Vision: Safe Operation of Autonomous Ground Vehicles in Dynamic Environments

A. Howard, L. H. Matthies, A. Huertas, M. Bajracharya and A. Rankin

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A. firstname.lastname@jpl.nasa.gov

Summary. This paper describes an integrated system for real-time detection and tracking of pedestrians from a moving vehicle. We use stereo vision as the primary sensor, and show that this sensor has a number of practical advantages over monocular vision. These include the ability to quickly identify regions-of-interest in an image, classify these regions based on shape, and track detected pedestrians in three-dimensional world coordinates. This system can also utilize monocular appearance-based algorithms, using regions-of-interest with known scale to increase speed and reliability. Finally, stereo allows us to construct fast algorithms for detecting independent motion in a scene.

1 Introduction

A key barrier to the deployment of autonomous ground vehicles in urban environments is the ability of those vehicle to operate safely in the presence of moving objects such as pedestrians and other vehicles. Thus, a key research problem must be the development of sensors and algorithms capable of detecting and tracking movers in real-time and at useful ranges, such that the vehicle autonomy system has sufficient time to react appropriately. This problem brings a host of challenges: the environment is complex and three-dimensional, movers may be fully or partially occluded, and one must recognize potential movers even when they are stationary. This last point is crucial: it is not sufficient to simply detect motion; one must be able to detect a stationary pedestrian, for example, and anticipate that the pedestrian may step out onto the road.

In this paper, we consider the specific problem of real-time detection and tracking of pedestrians; to keep the problem tractable, we restrict ourselves to humans with upright postures (e.g., standing, walking or running). Our approach relies on stereo vision as the primary sensor, which has a number of advantages. First, stereo provides dense 3D range data with high angular resolution, allowing us to both recognize pedestrians and correctly place

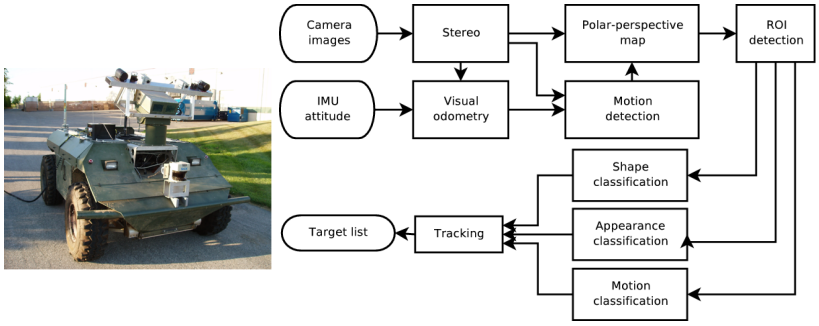

Fig. 1. (a) JPL stereo vision system mounted on the GDRS XUV (inner camera pair on the pan-tilt unit). (b) Pedestrian detection system diagram.

them them in the world. Second, stereo can provide appearance data at a variety of wavelengths (e.g., visible RGB or IR), allowing the use of monocular appearance-based recognition techniques. Third and finally, stereo is an off-the-shelf technology with both software and hardware implementations (the latter capable of processing 1024x768 images at 10Hz) and one that is readily adaptable to different vehicles and requirements through the choice of camera resolution, lenses and baseline. Stereo also has at least one disadvantage when compared with other ranging sensors such as radar or ladar: while stereo has excellent angular resolution, the range resolution decays quadratically. Consequently, in this paper, we introduce a novel map representation called the polar-perspective map (see §3) that partially addresses this limitation.

The complete stereo vision based pedestrian detection system is described in the next section. This system is currently being developed for the ARL Robotics CTA Program, and is deployed on the General Dynamics Robotic Systems XUV vehicle shown in Figure 1. We believe that the general approach will also extend naturally to other problems in dynamic scene understanding.

2 A Stereo Vision System for Pedestrian Detection

The high-level architecture for stereo vision based pedestrian detection is shown in Figure 1(b). The key blocks are as follows.

- **Stereo vision.** The stereo vision module takes synchronized images from a pair of cameras and computes a dense range image. Stereo is by far the most time-consuming step in the end-to-end pedestrian detection pipeline, requiring approximately 100ms to process 512x384 images on a commodity CPU. To reduce this overhead, we are currently developing an FPGA stereo implementation that can process 1024x768 images at 10-15Hz with no significant CPU utilization.
- **Visual odometry.** The visual odometry module takes two pairs of stereo images and computes the frame-to-frame camera motion. In practice, thus

computation can be greatly accelerated by utilizing the intermediate data products from the stereo pipeline, such as the rectified and stereo disparity images. With this enhancement, typical processing times are less than 10ms for 512x384 images.

Note that visual odometry serves two purposes in this architecture. First, it can be used to estimate the pose of the camera by integrating frame-to-frame motion estimates (particularly useful on legged vehicles, traditional wheeled odometry is not available). Second, as described in §5, visual odometry aids the interpretation of dynamic scenes by allowing us to compensate for the camera motion. The stereo and visual odometry modules both rely on well tested, off-the-shelf technologies (see [3, 6, 7, 8, 9], for example) and are not discussed at length in this paper.

- **Region-of-interest (ROI) detection.** As a first step towards finding pedestrians, the ROI detector searches the scene for any vertical surfaces, including, but not limited to, upright humans. This module is described in §3, where we introduce the *polar-perspective map* to enable fast ROI detection at long ranges.
- **Shape-based classification.** Using range image from the stereo module, we can construct a 3D point cloud for every ROI, and classify this cloud according to the point distribution. In §4, we describe a learned shape-based classifier that distinguishes between pedestrians and non-pedestrians.
- **Appearance-based classification.** Pure shape based classification has some limitations; it is difficult, for example, to distinguish between a human and a similarly shaped tree or shrub. Existing monocular pedestrian detection algorithms (e.g., [10, 12]) can be applied to the RGB pixel data in each ROI, with the expectation that these algorithms will run faster and produce better results when applied to a ROI with known scale.
- **Motion-based classification.** Motion is another important cue for distinguishing between pedestrians and non-pedestrians. In §5 we describe method for detecting motion from a moving camera, using visual odometry and stereo range data to correct for the camera ego-motion.
- **Tracking.** The tracking module maintains a track for each detected pedestrian, with a position and velocity estimate (in vehicle or world coordinates) and the fused output of the various classifiers. The implementation of the tracker is relatively straight-forward, since the availability of range data allows us to use a conventional Kalman Filter with nearest-neighbor data association. Unlike monocular vision systems, the tracker need not get confused when one pedestrian passes in front of another.

The architecture described above is intended to be fairly general, such that it can encompass most existing monocular vision techniques, while leveraging stereo vision to improve speed and reliability. In the following sections, we will discuss basic algorithms in greater detail and present experimental results from a system installed on the GDRS XUV.

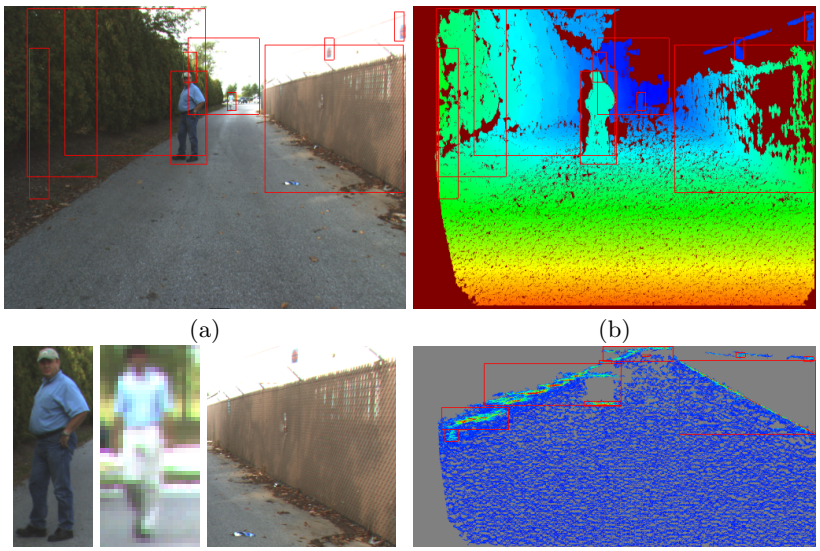

Fig. 2. ROI detection on the polar-perspective map. (a) Original RGB image with bounding boxes around the regions on interest. (b) False-color stereo disparity image: red pixels have high disparity (nearby points), blue pixels have low disparity (distant points). (c) Three of the ten regions of interest, including two humans at 7m and 37m, respectively. (d) Polar-perspective map of pixel count; cells with low pixel counts are shaded in blue, cells with high pixel count (vertical surfaces) are shaded in red. The large blue area corresponds to the ground plane, while the linear features on the left and right correspond to the tree-line and fence-line respectively. Bounding boxes for the detected ROI area also shown.

3 Fast ROI Detection on a Polar-Perspective Map

The first step in the pedestrian detection algorithm is finding regions of interest (ROI) in the image. To do this, we search the scene for vertical surfaces, thereby capturing both pedestrians (assumed to be upright) and non-pedestrian objects such as street signs, poles and trees. The detected ROI are passed on to the classifier stage described in §4 to sort pedestrians from non-pedestrians. The basic ROI detection algorithm is as follows:

1. Project stereo range data into a two-dimensional grid map centered on the camera, and accumulate statistics such as pixel count, mean elevation and elevation variance in each grid cell.
2. Segment the map into 'blobs' based on cell statistics. For example, a single blob may correspond to a group of connected cells whose elevation variance is greater than some threshold.
3. Back-project each blob into the original image to form a ROI. The ROI can be expressed as a bounding box or a list of individual pixels.

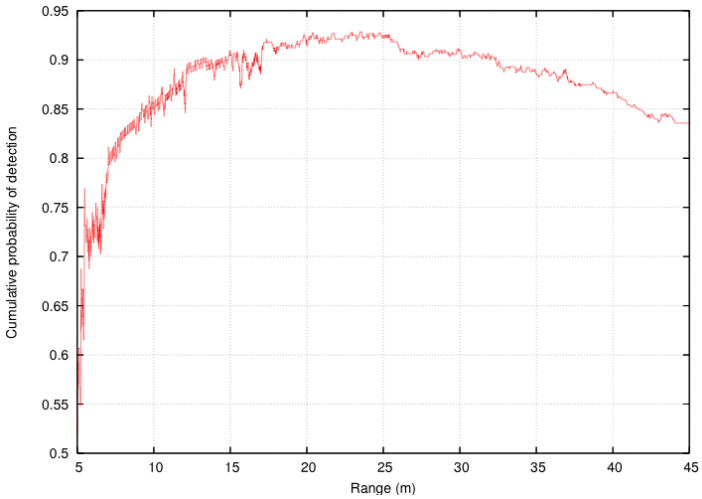

Fig. 3. ROI detection results, showing the cumulative probability of detection (PD) as a function of range. Results generated from a hand-labeled sequence of 500 stereo images, at 1024x768 resolution, 60° horizontal FOV and 0.50m baseline.

This algorithm is most reliable when the map is horizontal, i.e., when the map stays level with respect to gravity rather than pitching and rolling with the vehicle. Thus, in some applications, it may be useful to measure the camera attitude with respect to gravity and transform the stabilized stereo range points in the map frame. Camera attitude can be easily measured using an inertial measurement unit or an inexpensive set of accelerometers.

The finite angular and range resolution of the stereo sensor presents some challenges when attempting to detect ROI at longer ranges. On a regular Cartesian grid, the pixels making up a distant pedestrian may be distributed over a large number of cells, and segmentation algorithms looking for connected components will consequently fail. We have therefore developed a novel map representation that naturally captures the variable resolution of the stereo range data and preserves the utility of fast connected components algorithms. The *polar-perspective map* (PPM) employs a regular grid over a two dimensional space consisting of the bearing and inverse range to each point. Mathematically, the transformation between the polar-perspective space (d, θ) and the Cartesian space (x, y) is given by:

$$d = (x^2 + y^2)^{-\frac{1}{2}} \quad \theta = \tan^{-1}(y/x) \tag{1}$$

In this space, each grid cell subtends a constant angle, while the linear dimension varies with range (nearby cells are small, distant cells are large). Thus, unlike the Cartesian map, the resolution of the polar-perspective map can be tuned to exactly match that of the sensor.

Figure 2 illustrates the ROI detection process, showing the input images (color and disparity) and the polar-perspective map of total point count. Contiguous blobs are extracted from the PPM and back-projected into the original image to form the ROI (red boxes). In this particular example, the process has detected two ROI containing pedestrians (at ranges 7m and 37m) and a number of distracter objects (fences and vegetation). Quantitative results for ROI detection are shown in Figure 3. This plot shows the cumulative probability of detection as a function of the maximum range, indicating that there is a 90% probability of detection between ranges 0 and 25m, and an 85% probability of detection between ranges 0 and 45m. These results were generated using a hand-labeled training set of approximately 500 images, with 1024x768 resolution, 60° horizontal field-of-view and 0.50m baseline.

At long ranges, missed detections usually correspond to situations in which a pedestrian is standing close to another vertical object, such that the two merge together in the PPM. While these merged objects are usually detected as ROI, we do not count them as valid detections; the classifiers described in the next section can become confused when presented with multiple objects. At short ranges, missed detections usually correspond to situations in which pedestrians 'break up' into a number of ROI corresponding to separate body parts (e.g., one for each of the torso, arms and legs). That is, at close range, the PPM may have too much resolution. This suggests that the simple-but-fast blob detection algorithm described above should be augmented with a second stage that merges or splits detected ROI, possibly using a prior model.

Note that there are many possible variants on the polar-perspective map described here. At the simple end of the spectrum, for a camera that is horizontal to the ground, one can construct a basic map using the image column number and stereo disparity. At the complex end of the spectrum, we can adjust the grid spacing on both angular and radial axes such that the map smoothly transitions from a regular Cartesian grid in the near field to a polar-perspective grid in the far field. This latter approach may be advantageous if the algorithm must operate at both near and far ranges.

4 Classification using Shape and Appearance

Having extracted regions of interest from the image, the next step is to classify each ROI as being either a pedestrian or non-pedestrian. Given that we have access to range data from stereo, we can compute, for each pixel in the ROI, the corresponding 3D point. Figure 4, for example, shows some point clouds for pedestrians and non-pedestrians detected at different ranges. Several features should be noted: first, data along the depth axis is relatively uninformative, as range uncertainty quickly swamps the signal from the depth profile; second, since we assume the direction of the gravity vector is known, point clouds can always be oriented into an upright posture; and third, the effect of increasing range is to reduce the density of the point cloud, without changing its scale.

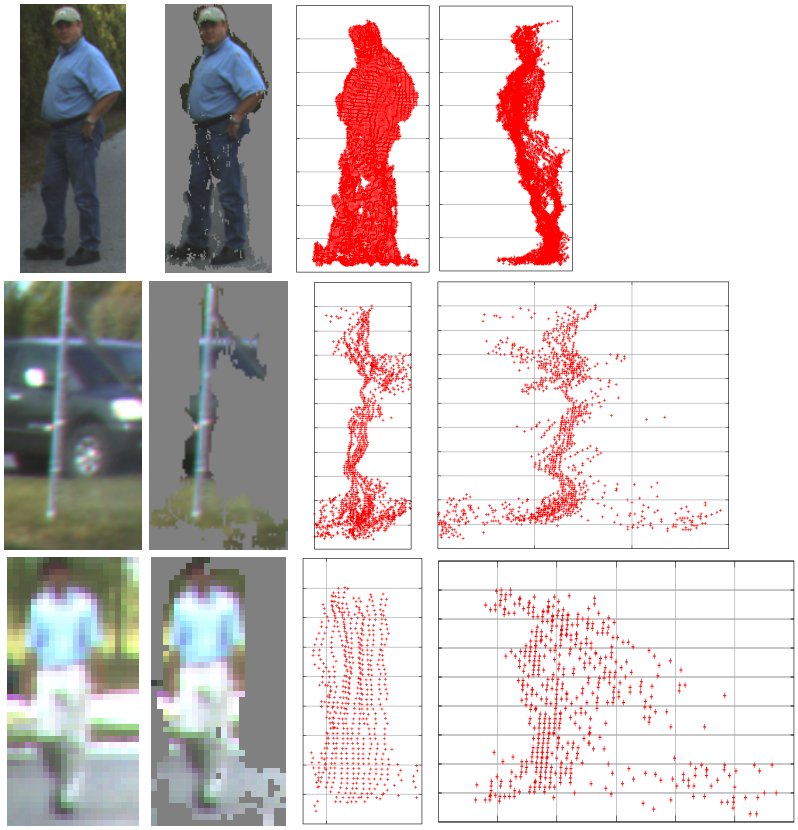

Fig. 4. Sample regions of interest for objects at 7m, 24m and 37m. (a) Sub-image corresponding to the ROI bounding box. (b) Sub-image after background removal (the foreground objects are somewhat fattened by the stereo correlation window (which is typically 7x7 or 9x9 pixels). (c) & (d) 2D projections of the stereo point cloud, both face-on and in-profile. The grid lines denote 1m intervals in the horizontal direction and 0.25m intervals in the vertical direction.

Consequently, one can construct a shape-based classifier that ignores the depth component and assumes that features are already scale and rotation invariant.

A simple classifier can be constructed by considering the central moments of the 2D point clouds. For example, in the two dimensional feature space consisting of the standard deviations σ_x and σ_y (corresponding to the width and height axis of the point cloud, respectively) one can show that humans are narrowly clustered around 0.15m and 0.50m. We have therefore implemented a naive Bayes classifier based on the first few central moments (up to order two), and trained this classifier against hand-labeled images (estimating the parameters of a pair of one-dimensional Gaussians for each feature). To classify

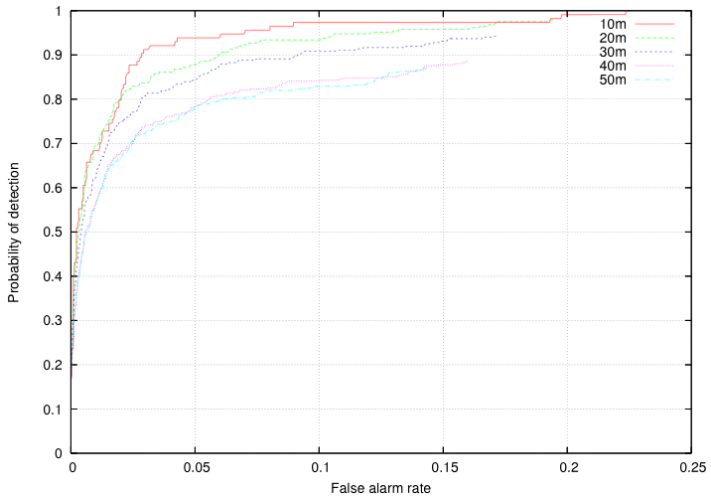

Fig. 5. Receiver Operating Characteristics curves (ROC) curves for a naive Bayes classifier using central moments. Each curve corresponds to a different maximum range cutoff.

a ROI, we compute the response from each feature Gaussian and combine using Bayes rule. Sample results are shown in Figure 5; this classifier was trained on 200 hand-labeled images, spanning ranges 5 to 50m, then tested against a second set of 300 labeled images covering similar ranges. Image resolution is 1024x768, with a stereo baseline of 0.50m. Note that different curves in this plot corresponds to different range cutoffs: for pedestrians within 20m, the classifier has a probability of detection (PD) of 80% with a false-alarm rate (FAR) of 2%; at 50m, these numbers are 75% and 3% respectively. Unsurprisingly, classifier performance degrades at longer ranges.

Much more sophisticated classifiers and feature sets are of course possible (e.g., support vector machines and mixture of Gaussians), and will likely improve on the results shown here (bearing in mind that performance is ultimately limited by the number of pixels in the ROI).

The key advantage of shape-based classifiers is that they are fast, simple to train and robust to illumination conditions and clothing type. On the other hand, the sample images in Figure 4 also demonstrate the need for appearance-based classification: at longer ranges, a pole and a pedestrian produce similar point clouds, such that a human observer would be hard pressed to distinguish between the two using shape alone. In contrast, it is not at all difficult to make the classification when presented with the ROI RGB images. There are of course many approaches to monocular appearance-based classification that could be applied this context, and we will not address these here. Instead, we restrict ourselves to making some practical observations. First, the ROI is more than just a bounding box on the image, since it contains a list of

all the pixels lying on the target of interest; this allows us to trivially mask out the background pixels, to produce the images shown in column (b) of Figure 4. For a learned classifier, background removal can significantly improve both the training and test error. Second, the shape-based classifier provides a simple mechanism for collecting training images for an appearance-based classifier. Thus we do not need to hand-label hundreds or thousands of images to generate a training set; instead, we can take the vehicle to an environment that is devoid of distracter objects (such as an empty parking lot) and quickly collect a set of positive examples. Conversely, we can drive the vehicle through a cluttered environment that is devoid of pedestrians to collect a set of negative examples.

5 Motion-based Classification: Detecting Motion from a Moving Vehicle

Motion can be used as a cue for detecting and classifying pedestrians. The main complication is, of course, the motion of the camera: simple image differencing techniques will pick out the entire image, not just the independent movers. We therefore adopt a two step approach in which we first estimate and correct for camera movement, and then apply a local change detector to pick out residual motion (similar to the approaches described in [1] and [11]). The algorithm is as follows:

1. Use visual odometry to determine the camera motion between frame A and frame B. Our VO algorithm (which is based on the approach described by Hirschmuller [6]) is very robust to outliers and ignores the independent movers in the scene.
2. Use the motion estimate and stereo range data to construct a predicted image at time B. That is, transform each point from frame A to frame B using the motion estimate, then back-project through the camera model to form the predicted image.
3. Compare the predicted and measured images at time B using a correlation-based optical flow algorithm. For each pixel in the measured image, we determine the local shift that will maximize the correlation with the predicted image; pixels with non-zero shift indicate independent motion in the scene.

The second step of this algorithm is illustrated in Figure 6: the predicted and measured images are very similar, despite the fact that they are taken 1.4 meters apart (these images have been pre-filtered with a Laplacian of Gaussians to improve the correlation search). Figure 6 also has a set of close-ups from the original image, showing a pedestrian crossing the road at a distance of 50m, clearly picked out by the motion detection algorithm.

In practice, the effectiveness of this algorithm is limited by the quality of the stereo range data, since any errors in range will be detected as motion

Fig. 6. Motion detection from a moving vehicle. Top row: raw images from a moving vehicle, taken 0.4 seconds and 1.4 meters apart. Middle row: predicted and measured images after motion compensation (with Laplacian-of-Gaussian filtering); note the movement of the pedestrian in the foreground right, in images that are otherwise very similar. Bottom row: close-up of a pedestrian crossing the road in the background center, approximately 50m; the output of the motion detector is shown in red.

(particularly around occluding boundaries). Some post-filtering of the optical flow is therefore required to eliminate spurious detections. It should also be noted that motion along the optical axis is difficult to detect, particularly at long distances, since it produces little if any change in the image. Thus, in the scene shown in Figure 6, the algorithm generates a strong response for pedestrians crossing the road, but a weak response for people on the sidewalk.

Interestingly, with this approach, it may be possible to detect certain motions beyond the effective range of stereo (e.g., in pixels with zero disparity).

Although these detections cannot be placed in a map, they can be used as a cue to either slow the vehicle or aim a high-resolution sensor at the target.

6 Related Work

The approach described in this paper bears much in common with the multi-cue pedestrian detection system recently proposed by Gavrila and Munder [4]. The PROTECTOR system uses sparse stereo vision to rapidly extract regions of interest, applies shape and texture based classifiers, then performs a final dense stereo validation. The authors also propose an interesting method for optimizing system parameters (such as identification thresholds) through systematic exploration of the ROC space (which considers both the probability of detection and the false positive rate).

Stereo vision systems for pedestrian detection are also described by Grubb [5], using an ROI detection plus classification approach, and by Bertozzi et al. [2] using IR imagery to extract warm bodies, sparse stereo with 'V-Space' ROI detection, and template-based head matching. It should be noted that the approach described in the current paper has also been applied, without modification, to IR stereo imagery. The motivation here is not to use a human's heat signature per-se, but rather to enable safe operation of autonomous vehicles under all lighting conditions (day and night).

7 Discussion and Conclusion

The analysis we have performed to date supports the proposition that stereo vision, as a sensor, is an excellent fit to the problem of pedestrian detection for safe vehicle operation. Stereo retains all the capabilities of monocular vision, has better angular resolution than currently available ladars or radars, has no moving parts or emitted radiation, and is amenable to low-power FPGA implementation. It should also be noted that while correlation-based stereo techniques have difficulty calculating range on some surfaces (such as narrow vegetation), these same techniques are extremely effective on upright human beings in a wide variety of clothing types (including camouflage). Compared with ladar, the main weakness of stereo vision is its relatively poor range resolution, the polar-perspective map described in §3 partially addresses this problem, however, and allows us to retain reasonable detection rates out to a range of about 50m, at a resolution of 1024x768, 60° FOV and 0.5m baseline.

Ultimately, the maximum range of this system is constrained by two factors: range resolution and pixels-on-target. Range resolution determines whether or not we can pick out distant figures from the background clutter (stereo disparity on pedestrians must be non-zero), while the number of pixels-on-target dictates the false alarm rate from the classification stage. Range resolution and pixels-on-target are in turn determined by the image

resolution, field-of-view and camera baseline. Since the field-of-view and baseline are generally dictated by engineering constraints (such as the location and number of cameras mounted on the vehicle), it with increased image resolution that we see the greatest opportunity to increase the maximum detection range. For example, we have recently begun experiments with a pair of commercial 16 mega-pixel cameras (4.8K by 3.2K pixels) that should theoretically improve the maximum range by a factor of five (to 250m). With such cameras, the key research problem lies in designing stereo algorithms that can operate in real-time, using techniques such multi-resolution processing and focus of attention.

Acknowledgments

This work was sponsored by the ARL Robotics Collaborative Technology Alliance (Robotics CTA), DARPA Learning Applied to Ground Robotics (LAGR) and DARPA Urban Grand Challenge (UGC) programs.

References

1. M. Agrawal, K. Konolige, and L. Iocchi. Real-time detection of independent motion using stereo. In *Motion and Video Computing Workshop (WACV/MOTION'05)*, 2005.
2. M. Bertozzi, E. Binelli, A. Broggi, and M. D. Rose. Stereo vision-based approaches for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR'05)*, 2005.
3. Y. Cheng and M. Maimone. Visual odometry for the mars exploration rovers. In *IEEE Robotics and Automation Magazine*, 2006.
4. D.M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1), 2007.
5. G. Grubb. *3D vision sensor for improved pedestrian safety.* PhD thesis, RSISE, Australian National University, 2005.
6. H. Hirschmuller, P.R. Innocent, and J.M. Garibaldi. Fast, unconstrained camera motion estimation from stereo without tracking and robust statistics. In *Int. Conf. on Control, Automation, Robotics and Vision (ICARCV'02)*, 2002.
7. L. Matthies. *Dynamic stereo vision.* PhD thesis, Department of Computer Science, Carnegie Mellon University, 1989. CMU-CS-89-195.
8. D. Nister, O. Naroditsky, and J. Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1), 2006.
9. C.F. Olson, L. H. Matthies, M. Schoppers, and M.W. Maimone. Stereo egomotion improvements for robust rover navigation. In *Int. Conf. on Robotics and Automation (ICRA'01)*, 2001.
10. Y. Ran, I. Weiss, Q. Zheng, and L. S. Davis. Pedestrian detection via periodic motion analysis. *International Journal of Computer Vision*, 71(2), 2007.
11. A. Talukder and L. Matthies. Real-time detection of moving objects from moving vehicles using dense stereo and optical flow. In *Int. Conf. on Intelligent Robots and Systems (IROS'04)*, 2004.
12. P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2), 2005.