Proceedings of the IEEE ICRA 2009
Workshop on People Detection and Tracking
Kobe, Japan, May 2009

Results from a Real-time Stereo-based Pedestrian Detection System on a Moving Vehicle

Max Bajracharya, Baback Moghaddam, Andrew Howard, Shane Brennan, Larry H. Matthies

***Abstract*— This paper describes performance results from a real-time system for detecting, localizing, and tracking pedestrians from a moving vehicle. The end-to-end system runs at 5Hz on 1024x768 imagery using standard hardware, and has been integrated and tested on multiple ground vehicles and environments. We show performance on a diverse set of ground-truthed datasets in outdoor environments with varying degrees of pedestrian density and clutter. The system can reliably detect upright pedestrians to a range of 40m in lightly cluttered urban environments. In highly cluttered urban environments, the detection rates are on par with state-of-the-art non-real-time systems [1].**

I. INTRODUCTION

The ability for autonomous vehicles to detect and predict the motion of pedestrians or personnel in their vicinity is critical to ensure that the vehicles operate safely around people. Vehicles must be able to detect people in urban and cross-country environments, including flat, uneven and multi-level terrain, with widely varying degrees of clutter, occlusion, and illumination (and ultimately for operating day or night, in all weather, and in the presence of atmospheric obscurants). To support high-speed driving, detection must be reliable to a range of 100m. The ability to detect pedestrians from a moving vehicle in a cluttered, dynamic urban environments is also applicable to automatic driver-assistance systems or smaller autonomous robots navigating in environments such as a sidewalk or marketplace.

This paper describes results from a fully integrated real-time system capable of reliably detecting, localizing, and tracking upright (stationary, walking, or running) human adults at a range out to 40m from a moving platform. Our approach uses imagery and dense range data from stereo cameras for the detection, tracking, and velocity estimation of pedestrians. The end-to-end system runs at 5Hz on 1024x768 imagery on a standard 2.4GHz Intel Core 2 Quad processor. The ability to process this high resolution imagery enables the system to achieve better performance at long range compared to other state-of-the-art implementations. Because the system segments and classifies people based on stereo range data, it is largely invariant to the variability of pedestrians' appearance (due to different types and styles of clothing) and scale. The system also handles different viewpoints (frontal vs. side views) and poses (including articulations and walking) of pedestrians, and is robust to objects being carried or worn by them. Furthermore, the system makes no assumption of a ground-plane to detect or track people, and similarly makes no assumption about the predictability of a person's motion other than a maximum velocity.

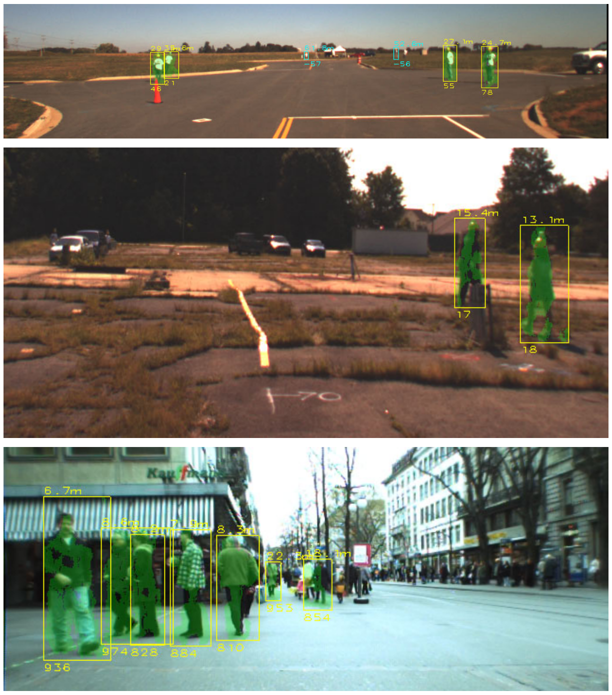


Fig. 1. Examples of test scenarios and the output of our pedestrian detection system (yellow boxes are detections with range and track ID text and a green overlay of the segmented person; the cyan boxes are missed detections).

The performance of the system is demonstrated on a variety of ground-truthed datasets in various outdoor environments, with different degrees of person density and clutter. An example of these scenes is shown in Figure 1. The majority of datasets used to evaluate the system consist of scenarios simulating the operation of an unmanned ground vehicle (UGV) traveling at moderate speed in semi-urban terrain (paved roads with light clutter and people walking along or into the road). In these scenarios, the system is capable of initial detections of pedestrians up to 60m, and reliable detection and tracking of pedestrians up to 40m, which correspond respectively to 30 pixel and 45 pixel tall pedestrians for our cameras. We also present performance results of our system on recently published datasets of crowded street scenes. Although not specifically designed for

The research described in this publication was carried out at the Jet Propulsion Laboratory, California Institute of Technology, with funding from the Army Research Lab (ARL) under the Robotics Collaborative Technology Alliance (RCTA) through an agreement with NASA

All authors are with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109

highly cluttered urban environments, we show that results of our real-time system are comparable to the state-of-the-art systems that are designed to operate in these environments.

II. Related Work

There has been extensive research on pedestrian detection from manned and unmanned ground vehicles using scanning laser rangefinders (LIDAR) and monocular and stereo vision in visible, near infrared, and thermal infrared wavelengths. Most such work assumes the scene contains a dominant ground plane that supports all of the pedestrians in upright postures. Maximum detection ranges tend to be 30m or less. Rates of missed detections and false alarms are not good enough to be satisfactory in deployed systems. Most prior work on pedestrian detection has been done for applications to smart automobiles, robotic vehicles, or surveillance. This literature is very large, so we only cover recent highlights and main trends here.

Research on pedestrian detection for smart automobiles has employed monocular vision [2], [3], [4] stereo vision [5], [6], [7], [8], [9] and LIDAR [10]. Vision-based methods have used visible [2], [3], near infrared [4], and thermal imagery [8]. Most work in this area has been strongly motivated by the requirement to be very low cost in eventual production. The approaches generally follow the architecture of detecting regions of interst (ROIs), classifying these regions, and tracking them.

Work on pedestrian detection for robotic vehicles in outdoor applications [11], [12], [13], [14], [15] includes methods that do range sensing with 2D LIDAR, 3D LIDAR, stereo vision, and/or structure from motion and do image sensing with visible and/or thermal infrared cameras. At a high level, algorithm architectures are analogous to the systems for the automotive domain, involving ROI detection, classification, and tracking, though the order and details of these steps differ. As a group, there is more emphasis in this domain on classification based on the 3D shape of the objects as perceived by LIDAR or stereo vision than there is in the automotive domain. The feature extraction and classification algorithms tend to be simpler than those used in either the automotive or video surveillance domains. Several of these approaches have been tested as part of third party field experiments, with results discussed by Bodt [16].

Finally, work on pedestrian detection in the surveillance arena largely divides into work with image sequences from stationary cameras, where background subtraction and/or image differencing is used to detect moving objects [17], [18], and work that applies trained pattern classifiers to individual images [19], [20], [21], [22], [23]. The former group is less relevant here, because background subtraction and temporal image differencing are more difficult to use from moving cameras. The latter group uses a variety of feature extraction and classification methods to achieve better detection and false alarm rates than single-frame results reported in the automotive pedestrian detection literature; however, the results are not directly comparable because computational requirements are generally higher, the testing protocol often uses image databases where positive examples are already centered in image chips or does exhaustive search over position and scale of ROIs in test imagery, and because only individual frames are considered, the systems do not include any tracking.

III. System Description

Our system is fully described in earlier work [14], but we briefly summarize our approach here. We focus on two differences from our prior system: a slightly reduced feature set, and an improved tracker. Our system consists of the following steps:

- **Stereo vision** takes synchronized images from a pair of cameras and computes a dense range image.
- **Region-of-interest (ROI) detection** projects stereo data into a polar-perspective map and then segments the map to produce clusters of pixels corresponding to upright objects.
- **Classification** computes geometric features of the 3D point cloud of each ROI and classifies the object, resulting in a probability of being human.
- **Tracking** associates ROIs in sequential frames, accounting for vehicle motion, and estimates the velocity of the detected objects.

The system architecture allows the possibility of using appearance and motion features to improve the classification of people, but we currently do not make use of these features.

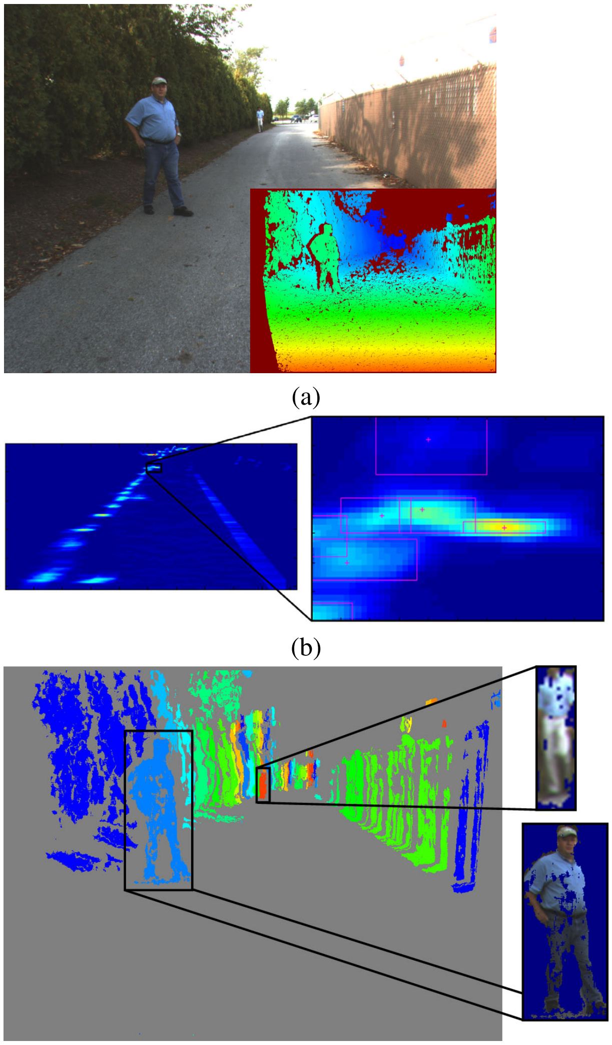
A. Stereo Vision

The first step in our system is to compute dense range data from stereo images. We use a multi-processor version of the real-time algorithm described by Goldberg [24] previously used on the NASA Mars Exploration Rovers and in the DARPA PerceptOR program. On a 2.4GHz Intel Core 2 Quad processor, the algorithm can process 1024x768 imagery at 10 frames/sec.

B. Region-of-Interest Detection

Detecting region-of-interest (ROI) areas from the stereo data serves as a focus-of-attention mechanism to reduce the runtime of subsequent classifiers and segments foreground pixels from background pixels in a region. This allows a shape-based classifier to be run on the 3D points that make up a specific object, rather than sliding a window over the image and explicitly performing foreground/background segmentation in each window.

The stereo range data is transformed into a gravity-leveled frame, accounting for the roll and pitch of the vehicle, and then projected into a two-dimensional polar-perspective grid map (PPM). The map is then segmented based on map cell statistics. Unlike a traditional Cartesian map, which is divided into cells of fixed size in Cartesian (x,y) space, the PPM is divided into cells with a fixed angular resolution but variable range resolution in polar (r, θ) space in order to preserve the coherency of the stereo range data. The PPM accumulates the number of range points projected into each cell. The map is then smoothed with an averaging filter with an adaptive bandwidth in polar space corresponding to a fixed bandwidth in Cartesian space. For computational efficiency the filter is implemented using an integral image of the map. After smoothing, the map gradient is used to find all of the peaks in the map, which are then grown to the inflection points in the gradients, resulting in a segmentation of the map. Because the minimum expected size of the objects being detecting is known, segmented blobs whose peaks fall within half of this size are then merged together. Figure 2 provides an example of a filtered PPM with ROI detections.


(a)

(b)

(c)

Fig. 2. An example of the stereo-based segmentation for region-of-interest detection. (a) shows the left image of a stereo pair with the resulting depth map (inset); (b) shows the polar-perspective map of point counts smoothed with an averaging filter with a close up of the map with segmented regions overlaid; and (c) shows the segmented regions in different colors, with examples of the foreground/background separation.

C. *Classification*

Geometric features of each segmented 3D point cloud are used to classify them as human or not human based on shape. After segmentation, a scene may contain hundreds of regions. To reduce the number of regions that must be classified, we first prefilter the regions with a fixed threshold on the width, height, and depth variance of each segmented region. This threshold is simply selected as the 3σ values obtained from the training data. After prefiltering, the features used for classification are computed for each region's point cloud.

We then compute geometry-based features for the remaining regions, including 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. The logarithmic and empirical logit transforms of these moments and counts are used to improve the normality of the feature distribution.

To compute the features, we start by centering the point cloud about the x-axis by its mean value and setting the minimum depth z and height y to zero. The first feature is defined by the logarithm of the 2^{nd} order moment of the height:

$$f_1 = -\log(\sigma_y^2) \tag{1}$$

The "soft-count" features are defined by the number of points that fall inside certain preset coordinate bounds (or volumes). Such count-based features ignore "true shape" and focus instead on the object's size or extent. Unlike moment-based features, count-based features are more tolerant of outlier noise and some artifacts of stereo processing. Naturally there are strong correlations between these two different sets of features. However, this correlation or redundancy can be quite helpful for modeling purposes. For the total number of points n in a blob point cloud, we define $n_x = \#(|x| < 1)$ as the number (subset) of 3D points whose x value is less than 1m (in absolute value), $n_{y_0} = \#(y < 2)$ and $n_{y_1} = \#(y > 1)$ as the number of points whose height value is less than 2m and greater than 1m, and $n_{z_0} = \#(z < 4)$ and $n_{z_1} = \#(z < 3.5)$ as the number of points with a depth value less than 4m and 3.5m respectively. We also define n_v to be the number of 3D points that satisfy all three width, height, and depth constraints simultaneously (i.e., the number of points that fall within the prescribed rectangular volume of size 1m x 2m x 4m). Although these constraints were selected empirically, the process could easily be automated. In order to normalize the data as well as account for uncertainty due to the sample size (n), we use a logit transform with an empirical prior count c:

$$f_2 = \log \frac{n_x + c_x}{n - n_x + c_x} \qquad f_3 = \log \frac{n_{y_0} + c_{y_0}}{n - n_{y_0} + c_{y_0}} \tag{2}$$

$$f_4 = \log \frac{n_{z_0} + c_{z_0}}{n - n_{z_0} + c_{z_0}} \qquad f_5 = \log \frac{n_v + c_v}{n - n_v + c_v} \tag{3}$$

$$f_6 = \log \frac{n_{y_1} + c_{y_1}}{n - n_{y_1} + c_{y_1}} \qquad f_7 = \log \frac{n_{z_1} + c_{z_1}}{n - n_{z_1} + c_{z_1}} \tag{4}$$

The rotationally-invariant features are the logarithms of the eigenvalues of the point cloud's covariance (inertia) matrix, where $(\lambda_x, \lambda_y, \lambda_z)$ correspond to the major, intermediate, and minor axes, respectively:

$$f_8 = -\log(\lambda_x) \quad f_9 = -\log(\lambda_y) \quad f_{10} = -\log(\lambda_z) \tag{5}$$

We note that f_8 would be redundant with f_1 if all the blobs were oriented correctly (upright and "facing" downrange). However, this is often not the case, due to artifacts in stereo processing, and especially at long ranges where blob point-clouds are often tilted and/or slanted.

Analysis of the shape features indicated that a linear classifier (with a linear decision boundary) was too simple to always work effectively. However, a more complex decision boundary can be achieved while still using a linear classifier (which is desirable for its computational efficiency and robustness) by expanding the feature set to use higher-order terms. Specifically, a quadratic decision boundary is modeled using the augmented feature set:

$$\mathbf{x} = [\ 1 \ \ \{f_i\} \ \ \{f_i f_j\}_{i<j} \ \ \{f_i^2\}\ ]^T \tag{6}$$

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.

D. Tracking

Tracking ROIs in the scene is used to both reduce incorrect detections and estimate the velocity of the detected objects. By associating ROIs across multiple frames, the single frame classifications can be aggregated to eliminate false positives. Similarly, using the positions of a tracked object from stereo and the motion of the vehicle, estimated by visual odometry [25] or provided by an inertial navigation system (INS), the velocity of the object can be computed and extrapolated to provide a predicted motion to a path planner. The tracking algorithm is designed to be extremely computationally efficient and makes very few assumptions about the motions of objects.

Tracking is implemented as the association of ROIs in sequential frames. The ROIs extracted in a new frame are matched to existing nearby tracks by computing a cost based on each ROI's segmented foreground appearance and then solving a one-to-one assignment problem. For computational efficiency and simplicity, the cost between an ROI and a track is computed by comparing the new ROI to the last ROI in the track. Only ROIs within a fixed distance are considered; the distance is computed by using an assumed maximum velocity of 2m/s in any direction for each object. The cost between ROIs is then computed as the Bhattacharyya distance of a color (RGB) histogram between each ROI. For computational efficiency, we solve the assignment problem with co-occurring minima. If an ROI does not match an existing track, a new track is started. Tracks that are not matched for a fixed number of frames are removed. To eliminate the incorrect detections that lead to false positives while still maintaining detections on true positives where the classification score dropped for a small number of frames, we temporally filter the scores with the median of three consecutive scores and require three consecutive frames of detection before making a classification decision. The velocity of tracks is estimated by fitting a linear motion model to the track. We estimate the position and velocity uncertainty by combining the expected stereo error with the model fit.

IV. EXPERIMENTAL RESULTS

The end-to-end system has been tested on datasets with hand-labeled ground-truth and integrated onboard a vehicle for live testing. The primary datasets were collected from the vehicle on which the system was integrated in semi-urban, lightly cluttered scenarios. The results on these datasets show that our system can achieve initial detections at a range of 60m, with detections reliable enough for autonomous navigation out to 40m. To demonstrate that the system's performance is competitive with state-of-the-art systems in highly cluttered, urban scenarios, we also make use of datasets published by Ess [1], [26]. We show that we can achieve performance similar to Ess on these datasets while running at real-time rates.

A. Semi-Urban Datasets

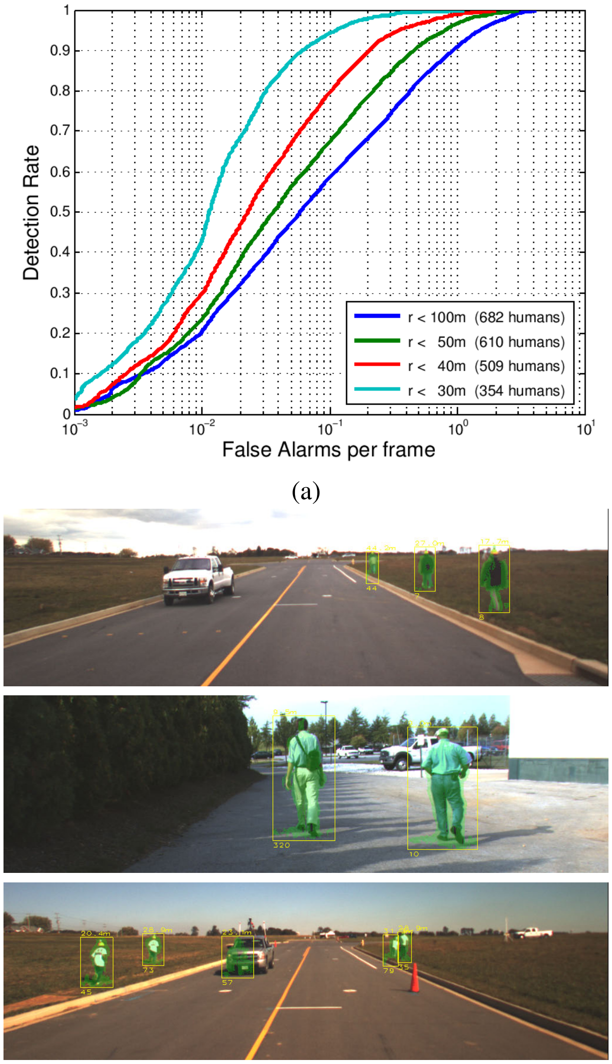
The primary datasets used to evaluate the system use input imagery from a 3 CCD color stereo camera pair with 1024x768 pixels, a 50 cm baseline, a field of view approximately 60 degrees wide, and with frame rates between 3.5Hz and 10Hz. The cameras were either mounted on the roof of an SUV at a height of approximately 2m above the ground, and pointed down by approximately 5 degrees, or on the pan-tilt head of an unmanned vehicle at a height of approximate 2m above the ground, and pointed down by 20 degrees. The scenarios include the vehicle driving down a road at speeds varying from 15 to 30 kph, with stationary mannequins and people standing, walking, and running along the side of and across the road in varying directions. The scene also contains stationary and moving cars, trucks, and trailers, along with stationary crates, cones, barrels, sticks, and other similar objects. In many cases, the pedestrians experience a period of partial to full occlusion by these objects or each other. Several variations of the scenario also include one or two people walking in front of the vehicle, weaving between each other and occasionally going out of the field of view.

The imagery was manually ground-truthed by annotating a bounding box around each person in the left image of each frame, to a range of approximately 100m. In total, our corpus includes approximately 6,000 annotated frames with approximately 10,000 annotated people, although we restrict our analysis to specific datasets which are representative of operational scenarios. Although people are annotated regardless of their posture or degree of occlusion, we only consider people who are in an upright posture with less than 50% occlusion for our analysis. We use the measure of the area of the intersection over the area of the union of the annotated and detected bounding boxes to declare a correct detection. However, for these datasets, we found that relaxing the common evaluation criteria of 50% intersection-over-union to 25% produced more meaningful results. This is because we are interested in detection at relatively long range where the segmentation error is dominated by the foreground fattening effect of stereo matching. Because the scenes are relatively uncluttered, using a looser matching criteria still remains representative of actual detections. In order to present results that are meaningful when developing a complete, autonomous system capable of safe navigation, we present our results as the probability of detection (Pd), defined as the number of detections divided by the true number of people in the scene, versus the false alarms per frame (FAPF), defined as the number of incorrect detections divided by the number of frames in the dataset. To illustrate the performance as a function of range, we restrict the detections and annotations to a maximum range.

To demonstrate the effectiveness of our feature set and classifier, we first present results on a cross-validation test over many of our datasets. Figure 3 (a) shows the performance of the system as an average of 1000 trials on a dataset combined from many different scenarios, totaling 4,396 frames with 3,409 annotated people. From these sequences, 21,824 ROIs were extracted and each curve was generated by randomly selecting 80% of these ROIs for training and using the remaining 20% for testing. The resulting number of effective frames in each test sequence is thus 879, and the average number of humans is shown in the plot for the respective range restriction. For this test, no temporal filtering was used to adjust the classification scores. Figure 3 (b) shows a sample of the images of the sequences used. The detections shown are indicative of the performance of the system (but are, in fact, based on a system trained without that sequence). Across our datasets, the system can achieve a 95% Pd at 0.1 FAPF for people less than 30m and 85% Pd at 0.1 FAPF for people less than 40m. For people out to 50m and 100m, the system achieves 95% and 90% Pd respectively at 1 FAPF.

Because the cross-validation results sample across all of the datasets being tested on, they do not necessarily provide compelling evidence that the system is effective in new, unseen scenarios. To demonstrate that our system is robust in new environments, we show the performance on individual sequences that have never been used for training. Although less statistically significant, they are perhaps more indicative of the performance to be expected of the fielded system. Figure 4 (a) and (b) show the results of the system without temporal filtering on two sequences held out from the training data. The same system was run on both datasets with no modification. As the plots show, the sequence shown in Figure 4 (a) is more difficult than (b), containing more clutter and occlusion. The system achieves well above 95% Pd at 0.1 FAPF for pedestrians less than 30m and 80% Pd for less than 40m. For a fielded system, we generally run at an operating point closer to 0.02 FAPF, which results in 90% Pd for <30m and 65% Pd for <40m, and maintain some degree of persistence of detected objects, propagating them with their predicted velocity for path planning.

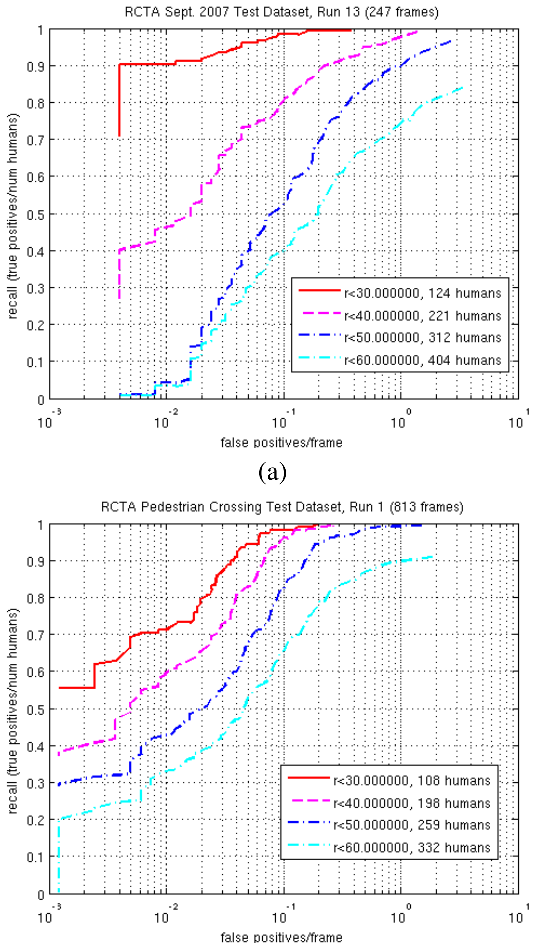
The main source of false alarms of our system in these environments is due to the over segmentation of vehicles. An example of a false alarm on the front of a pickup truck is shown in the lower image of Figure 3 (b). The individual distracter objects, such as barrels, tripods, and sign posts are only occasionally misclassified because they are normally segmented correctly. The main source of missed detections is due to variability of the stereo range data at long range, partial occlusion, and occasionally due to imprecise localization of the person due to under or over segmentation. Our system has some robustness to partial occlusion, but tends to break down after greater than 50% occlusion. The sequence shown in Figure 5 shows several examples of performance on occluding and overlapping people. The people in the near field are detected when they are unoccluded, or only slightly occluded. They are not detected when partially occluded either vertically (due to crossing the other person)



(a)

(b)

Fig. 3. (a) The performance resulting from 1000 trials of 80%/20% split cross-validation tests on 4,396 frames drawn from various scenarios. (b) Examples of images and detections from the various scenarios, with an example false alarm on the truck in the bottom image. The yellow boxes are detections, with a green overlay of the segmented person.



(a)

(b)

Fig. 4. The performance for two testing runs including people walking along and in the street, with moving cars and stationary distractor objects.

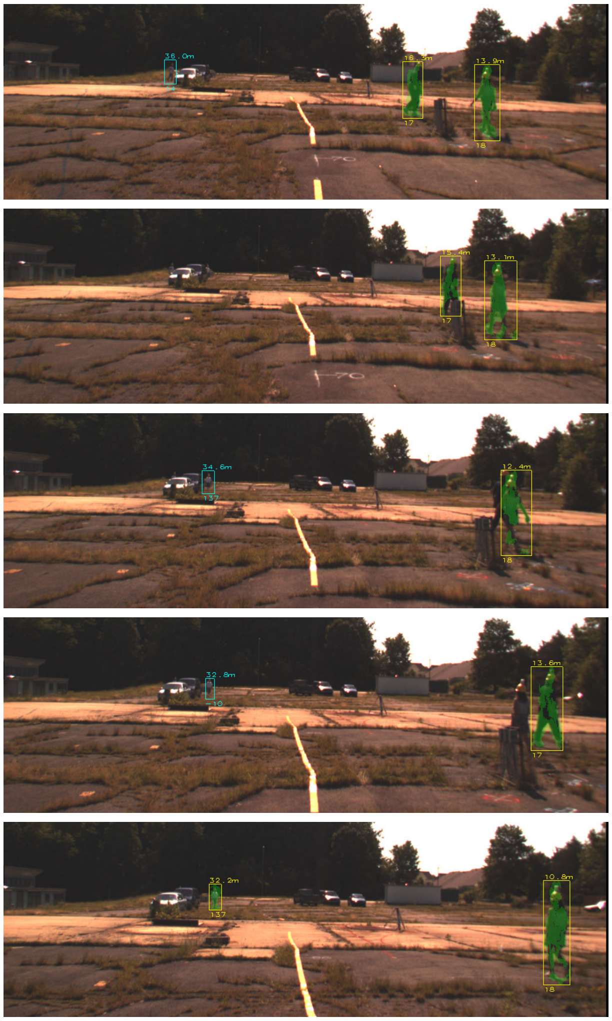


Fig. 5. A sequence of frames showing detections (yellow boxes, with green overlay the segmented person) and misses (cyan boxes) for people under occlusion. The number above the boxes indicates the range, and the number below indicates the track ID.

or horizontally (due to the posts). Notice, however, that the people are all tracked throughout the sequence (although with one incorrect association). The people in the far field are similarly not detected when they are partially occluded by the vehicles (or too far away), but are detected when they emerge into the open. The failure to detect partially occluded people is understandable because we only train a single classifier with data that does not contain many occluded people.

In addition to testing on ground-truthed datasets, the end-to-end system has been integrated into several systems for live testing. An earlier version of the system was fielded as part of the RCTA program "Safe Operations" test, as reported in [16]. The system described here has been integrated onboard the test vehicle for an upcoming test, for which results will be published in the future. The system has also been used to demonstrate autonomous navigation in a lightly cluttered dynamic environment on a small vehicle (with cameras at approximately 1m high and with a 12cm baseline) traveling at approximately 1m/s.

B. *Urban Datasets*

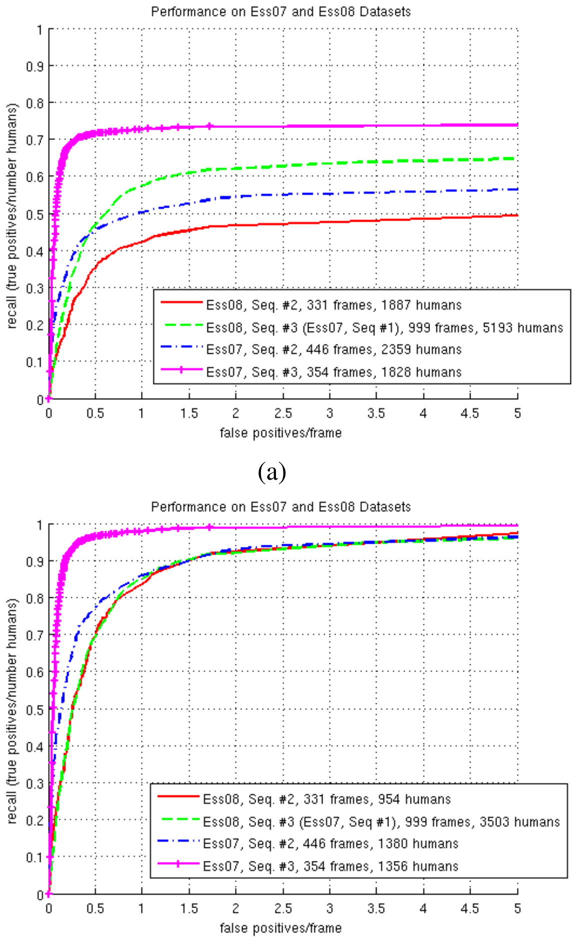
To illustrate that our system is competitive with other state-of-the-art stereo-based pedestrian detection systems, we also evaluated our system on datasets published by Ess [1], [26]. These datasets consist of 640x480 resolution color Bayer tiled imagery, taken at 15Hz, with a 40cm baseline camera pair pointed straight out at a height of approximately 1m. The scenarios are significantly more complex than the semi-urban data, with many people in a busy shopping district in Zürich, Switzerland, with significant occlusion, clutter, and motion. The annotations include all people whose torso is partially visible, and include children and partially upright postures, but not people sitting. To make a direct comparison to the results published by Ess, we use their detection criteria (50% intersection-over-union) and restrict the annotations used in the same way they do (with height greater than 80 pixels for sequence 2 of the 2008 data, and 60 pixels for all other data). We completely omit sequence 1 of the 2008 data because we were unable to generate acceptable stereo depth maps based on the camera models provided. The depth data density on all other sequences is acceptable, but not as dense as it could be, and results in reduced performance as discussed later. For direct comparison, we also train on exactly the same data as well (sequence 0 of the 2007 data).

The performance of our end-to-end system with the Ess test sequences using exactly the same evaluation criteria are shown in Figure 6 (a). Although the performance does not appear very good (between 0.4 and 0.7 recall at 1 false positive per frame, and with maximum achievable recalls between 0.5 and 0.75), it is very similar to the results reported by Ess. In fact, the results are slightly better at 1 FAPF on all sequences except sequence 2 of the 2008 data (which is due to less stereo coverage). Examples of the scenes, along with stereo and the predicted velocity of certain pedestrians, are shown in Figures 7 and 8. Notice that people are detected when they are in various poses or stages of walking and while carrying bags or briefcases. The main cause of the missed detections is simply due to a lack of stereo depth data density on people who are either too close or occluded. To illustrate this point, we also show the performance for the sequences where annotated people must have at least 10% stereo coverage (of the pixels defined by the annotated bounding box) in Figure 6 (b). Because our system relies on stereo data for both detection and classification, it can never find these people, nor would it be able to localize them to plan around them in a fully autonomous mode.

Our system misses detections and produces false positives in some understandable situations. For instance, it misses most children (left image of Figure 7), which were not included in any training data, and detects mannequins in shop windows or reflections of people in windows (right image of Figure 7). However, the majority of false detections is due to patchy stereo on flat surfaces such as buildings or cars, which results in the objects being over-segmented into a human sized objects (as seen on the car in the left image of Figure 8). Many times, this results in false positives high up on buildings (as seen in the center image of Figure 8), that could be removed by only considering people who might enter the street or be a danger. In other cases, explicitly detecting other objects such as cars would remove the false detections. Despite not designing for many of these situations, our system is capable of achieving competitive performance while running in real-time (10Hz on 640x480 imagery).

V. CONCLUSION

The results of our real-time, stereo-based pedestrian detection system show it to be effective at detecting people out to a range of 40m in semi-urban environments. It achieves results comparable with alternative approaches with other sensors, but offers the potential for long-term scalability to higher spatial resolution, smaller size, and lower cost than other sensors. It also performs similarly to state-of-the-art results from recent literature, while running at real-time rates.



(a)

(b)

Fig. 6. (a) The performance for sequences from [26] and [1] presented with the same evaluation criteria as their work. (b) the performance for the same sequences when all annotation that have less than 10% stereo coverage are eliminated, indicating that most of the misses in (a) are due to lack of stereo depth data on the people.

However, our system currently has some key limitations. Because the initial segmentation uses a projection into a 2D map, it cannot segment people or objects in close contact. To address this problem, we are investigating direct disparity-space and image-space segmentation techniques to provide regions of interest. Similarly, because we use a relatively small geometry-based feature set for classification, it is inherently limited. Any object with a similar shape to a person may be misclassified. To address this problem, we are investigating using appearance and motion features to improve classification. We are also using these extensions to handle the cases of pedestrians under partial occlusion and in non-upright postures.

REFERENCES

[1] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[2] A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: single frame classification and system level performance," in *IEEE Intelligent Vehicles Symposium*, 2004.

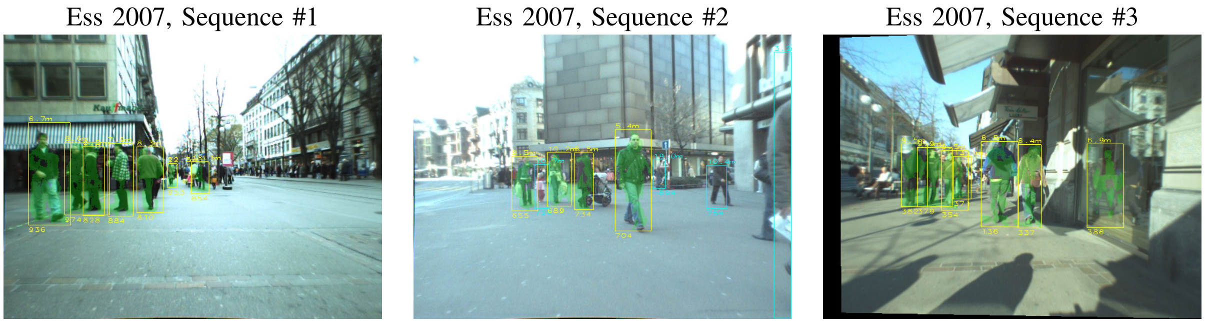


Fig. 7. Examples of detections (yellow boxes, with green overlay of segmented people) and misses (cyan boxes) for sequences from [26]. The false detection in the sequence 3 example is due to a reflection in the window.

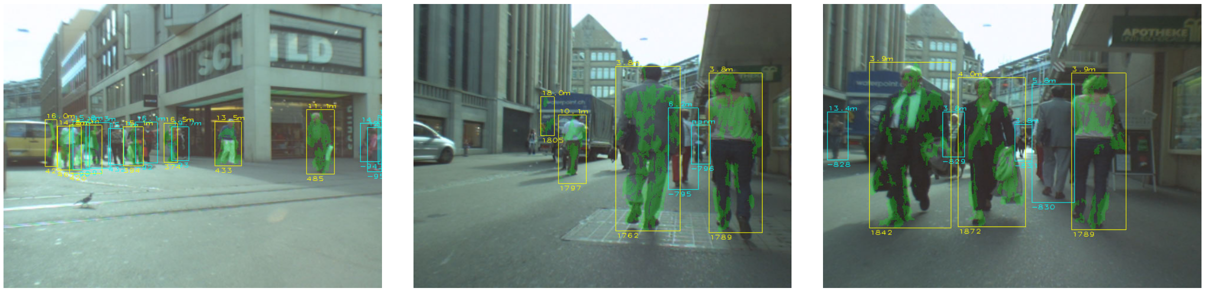


Fig. 8. Examples of detections (yellow boxes, with green overlay of segmented people) and misses (cyan boxes) for sequence 2 from [1]. There are false alarms on the car in the left image and the bus in the middle image. The misses are generally due to lack of stereo coverage or excessive clutter.

[3] G. Ma, S. Park, A. Ioffe, S. Muller-Schneiders, and A. Kummert, "A real time object detection approach applied to reliable pedestrian detection," in *IEEE Intelligent Vehicles Symposium*, 2007.

[4] R. Arndt, R. Schweiger, W. Ritter, D. Paulus, and O. Lohlein, "Detection and tracking of multiple pedestrians in automotive applications," in *IEEE Intelligent Vehicles Symposium*, 2007.

[5] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International Journal of Computer Vision*, vol. 73, no. 1, pp. 41–59, 2007.

[6] B. Liebe, N. Cornelis, K. Cornelis, and L. V. Gool, "Dynamic 3d scene analysis from a moving vehicle," in *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, 2007.

[7] M. Sotelo, I. Parra, D. Fernandez, and E. Naranjo, "Pedestrian detection using svm and multi-feature combination," in *IEEE Intelligent Transportation Systems Conference*, 2006.

[8] M. Bertozzi, A. Broggi, M. D. Rose, M. Felisa, A. Rakotomamonjy, and F. Suard, "A pedestrian detector using histograms of oriented gradients and a support vector machine classifier," in *IEEE Intelligent Transportation Systems Conference*, 2007.

[9] C. Tomiuc, S. Nedevschi, and M. M. Meinecke, "Pedestrian detection and classification based on 2d and 3d information for driving assistance systems," in *IEEE Intelligent Computer Communication and Processing Conference*, 2007.

[10] K. C. Fuerstenberg, K. Dietmayer, and V. Willhoeft, "Pedestrian recognition in urban traffic using a vehicle based multilayer laserscanner," in *IEEE Intelligent Vehicles Symposium*, 2002.

[11] S. M. Thornton, M. Hoffelder, and D. D. Morris, "Multi-sensor detection and tracking of humans for safe operations with unmanned ground vehicles," in *Workshop on Human Detection from Mobile Platforms, IEEE International Conference on Robotics and Automation (ICRA)*, 2008.

[12] L. E. Navarro-Serment, C. Mertz, and M. Hebert, "LADAR-based pedestrian detection and tracking," in *Workshop on Human Detection from Mobile Platforms, IEEE International Conference on Robotics and Automation (ICRA)*, 2008.

[13] A. Howard, L. Matthies, A. Huertas, M. Bajracharya, and A. Rankin, "Detecting pedestrians with stereo vision: safe operation of autonomous ground vehicles in dynamic environments," in *International Symposium of Robotics Research*, 2007.

[14] M. Bajracharya, B. Moghaddam, A. Howard, and L. Matthies, "Detecting personnel around UGVs using stereo vision," in *SPIE Unmanned Systems Technology X*, vol. 6962, March 2008.

[15] W. Abd-Almageed, M. Hussein, M. Abdelkader, and L. Davis, "Real-time human detection and tracking from mobile vehicles," in *IEEE Intelligent Transportation Systems Conference*, 2007.

[16] B. A. Bodt, "A formal experiment to assess pedestrian detection and tracking technology for unmanned ground systems," in *26th Army Science Conference*, December 2008.

[17] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *IEEE International Conference on Compuer Vision*, 2003.

[18] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," in *IEEE Trans. Pattern Analysis and Machine Intelligence (to appear)*, 2008.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[20] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[21] B. Wu and R. Nevatia, "Simultaneous object detection and segmentation by boosting local shape feature based classifier," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[22] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on Riemannian manifolds," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[23] E. Seeman, M. Fritz, and B. Schiele, "Towards robust pedestrian detection in crowded image sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[24] S. Goldberg, M. Maimone, and L. Matthies, "Stereo vision and rover navigation software for planetary exploration," in *IEEE Aerospace Conference*, March 2002.

[25] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *IEEE/RSJ Conf. on Intelligent Robots and Systems (IROS)*, 2008.

[26] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *International Conference on Computer Vision (ICCV)*, October 2007.