Multi-sensor Detection and Tracking of Humans for Safe Operations with Unmanned Ground Vehicles

Susan M. Thornton, *Member, IEEE,* Mike Hoffelder, and Daniel D. Morris, *Member, IEEE*

***Abstract*— This paper details an approach for the automatic detection and tracking of humans using multi-sensor modalities including 3D Ladar and long wave infrared (LWIR) video. By combining data from these sensors, we can detect individuals regardless of whether they are erect, crouched, prone, or partially occluded by other obstacles. Such algorithms are integral to the development and fielding of future "intelligent" unmanned ground vehicles (UGVs). In order for robots to be integrated effectively into small combat teams in the operational environment, the autonomous vehicles must maneuver safely among our troops and therefore must be capable of detecting stationary and moving people in cluttered scenes.**

I. Introduction

THE goal of this work is to robustly detect and track both stationary and moving humans from a moving unmanned ground vehicle (UGV) using data from both 3D Ladar and a long wave infrared (LWIR) camera. Detecting humans in any type of sensor data is a challenging problem due to the wide variety of positions and appearances which humans can assume. Stationary humans further complicate the task since motion cues can not be relied on to eliminate false alarms. In order to reduce false alarm rates, which can be significant in cluttered urban environments, algorithms are usually limited to detecting moving humans [2]-[3],[14] or to detecting upright humans in more ideal postures [4]-[8],[12]. These limiting assumptions are acceptable in the case of pedestrian detection for the automotive industry where the goal is to have an alert system that aids a human driver or in surveillance applications in which the camera is in a fixed location.

In the military context in which we are striving for a fully autonomous UGV that has 360 degree situational awareness, it is critical that we not only reliably detect upright people, but those that are prone on the ground (e.g. a hurt soldier) and those not moving. Since stationary humans have the potential to move at any time distinguishing them from other objects such as barrels and crates is important for effective and robust UGV path planning. To achieve this goal, we assert that it is necessary to fuse information from various sensor modalities. Unlike the fusion approach in [12], we have chosen Ladar and LWIR sensors which both provide day/night capability and operate under extensive environmental conditions.

In this paper, we first review a joint spatial-temporal solution to the Ladar data association problem [1] in which Ladar returns are considered as samples on an underlying world surface. This surface is explicitly modeled and then locally matched over time intervals. The result is a natural categorization of the world into stationary and moving objects. With a focus strictly on moving obstacles, the algorithm in [1] achieves a high rate of performance with the use of simple size features to distinguish between humans and other types of objects. We present improvements to the algorithm in [1] that focus on extracting more advanced shape-based features from the Ladar data and enable the detection of stationary humans.

Even with these advances, it is a challenge to differentiate a prone human in the Ladar from the massive amount of ground returns that are received, or even to detect upright humans when they are standing against walls or other large structures. In order to achieve these tasks, we use LWIR video to complement the capability of the Ladar. LWIR video provides information out to much longer ranges than the Ladar, and has the ability to highlight humans in a variety of non-ideal postures due to emissivity differences with their surroundings. We present a statistical and morphological approach to human detection in LWIR imagery. Many algorithms make the simplifying assumption that humans will always be hot compared to the surrounding environment [5], [14], but this is often not the case, so we discuss our approach to making the algorithm robust to such challenges.

In Section II, we summarize our method of object extraction and data association in Ladar data, and detail new features that have been implemented to improve performance. In addition, we present metrics and performance results on a baseline set of Ladar data. Section III contains a detailed presentation of our LWIR algorithm that extracts human regions-of-interest (ROIs). In Section IV, we briefly discuss our approach for fusing Ladar and LWIR at both the feature and detection levels. Section V summarizes our work.

Manuscript received January 15, 2008. This work was supported through collaborative participation in the Robotics Consortium which is sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0012

S. M. Thornton is with General Dynamics Robotic Systems, Pittsburgh, PA 15221 USA (phone: 412-473-2167; fax: 412-473-2190; e-mail: sthornton@gdrs.com).

M. Hoffelder and Daniel D. Morris are also with General Dynamics Robotic Systems, Pittsburgh, PA 15221 USA (e-mail: mhoffelder@gdrs.com, dmorris@gdrs.com).

II. LADAR ALGORITHM

A. Ladar Sensor Data

Our 3D data are the result of a pair of scanning Ladars that have been configured at fixed pan and tilt positions on the top of a sport utility vehicle (SUV). This dual configuration allows for nearly 180 degree field-of-view (FOV) in the direction the vehicle travels. In addition to the dual Ladars, the SUV has been equipped with an aluminum beam that holds a set of stereo LWIR video cameras, amongst a number of other sensor pairs, see Fig. 1.

Frames of data from the right and left Ladars are synchronized in time in order to account for either Ladar dropping frames or for unexpected differences in the frame rates of the two sensors. Each Ladar scans at approximately 10 Hz, providing a 2D grid based depth map for each scan. Although there is a small angular overlap of the two Ladars, we have initially chosen to do a simple concatenation of the depth maps from each sensor.

B. Object Extraction

In [1], an algorithm for detecting moving vehicles and people using a single scanning Ladar on a moving vehicle was presented. Since the data from the dual Ladar configuration appear to the algorithm as a single depth map, the approach is applied to the new dual data without modification. There are two key components to the approach: find objects in the scene and analyze their motion.

Object detection is achieved through a two-step process of eliminating ground returns and then implementing a contiguous region building technique that leverages the adjacency information in the angle-depth map created by the Ladars. This simple clustering approach has very low computational requirements, and works well for both large and small objects.

The ability to effectively isolate human objects depends on the performance of the ground removal algorithm. We found that modeling the ground with roughly horizontal planes [1] worked well at close range, but was unreliable at points far from the sensor and added undesirable computational complexity. The ground is now labeled by computing the elevation angle between neighboring points in the angle-depth map. Using the ground map in conjunction with the angle-depth map, a height-above-ground for each Ladar return is estimated. We use this height value to eliminate all points less than 0.25 meters above ground. This approach has the advantage of removing clutter due to tall grass and vegetation, as well as curbs in an urban setting, all of which pose difficulties to our algorithm if not eliminated.

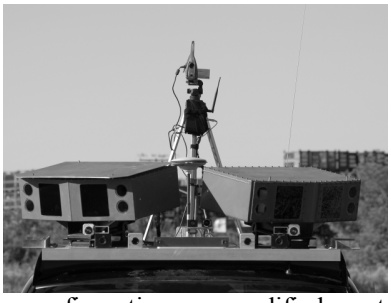

Fig. 1: Sensor configuration on a modified sport utility vehicle. The vehicle has 2 Ladars and a stereo pair of LWIR cameras.

C. Data Association and Tracking

We use a surface probability density model for 3D object registration [1]. The registration approach relies on explicitly modeling the object surface as a mixture of 3D Gaussians, $\rho_S\left({}^f X\right)$, centered at each sampled point:

$$\rho_S\left({}^f X\right) = \sum_i N\left(\boldsymbol{x}_i, \sigma_i^2\right) / n \,. \tag{1}$$

The covariances σ_i^2 are proportional to the sampling density, and hence to the distance from the Ladar. This models a wide variety of surfaces including coarsely sampled natural objects such as trees. Models are registered and scored by optimizing the Bhattacharya similarity measure which compares two density functions and gives an absolute similarity estimate enabling the goodness of a match to be assessed. The similarity measure is also useful for resolving matching ambiguities and detecting occlusions. A discrete implementation using convolution filtering enables real-time registration without being trapped by local minima.

Most of the work in moving object detection is achieved by clustering and registration. However, there are a number of sources of clutter as well as objects appearing and disappearing due to occlusions. These effects can lead to spurious motion estimates and hence false positives. Use of a Kalman filter tracker minimizes these effects, by enforcing motion consistency.

D. Classification Features

In order to detect stationary, as well as moving humans, we can not rely on simple size constraints for classification. The Ladar data provide significant shape information about an object, particularly at close range. We have implemented a new feature that quantifies this shape information and eliminates false alarms due to random clutter in the data. Random clutter refers to natural things in the environment such as thin vegetation, tall grass, or small branches which pose little danger to the UGV if in its path.

1) Shape-based feature extraction

Once points have been clustered into objects using our depth map region growing technique, the original 3D points from each human-sized object are projected onto a 2D plane, see Fig. 2. The projected cluster points are then used to create a 32 x 16 binary template which is aligned with the major axis of the cluster. As a measure of how uniformly distributed the returns are across the 2D grid, we compute a feature that we refer to as the fill-factor,

$$ff = 1 - \frac{\left(\#\,\text{empty bins}\right)}{\left(\text{total}\,\#\,\text{bins}\right)}. \tag{2}$$

Empirical analysis indicates that 2D binary templates of true humans, as well as other man-made objects, will be

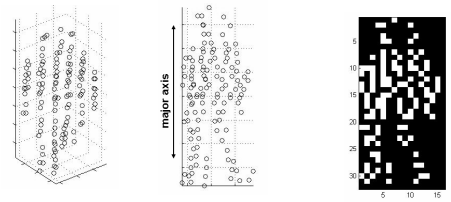

Fig. 2: (left) 3D Ladar points from one frame for a walking human; (center) Projection onto a 2D plane; (right) Binary 32x16 map. (*ff* = 0.2246)

roughly uniformly distributed, while a template resulting from random clutter which happens to be human size will be oddly shaped with points concentrated in small sections of the binary map. Random clutter clusters often meet human size constraints when several small clusters are erroneously grouped together (e.g. small clumps of grass and vegetation in close proximity to one another). The fill-factor value for these non-uniformly distributed clusters is much smaller.

Since the number of Ladar points associated with a cluster decreases dramatically with distance, the fill-factor value for humans at long range also decreases. Fig. 3 illustrates the returns for a human at *x* meters from the sensor compared to those at *2x* meters. As a result of the reduced number of returns at the longer distance, the fill-factor value is likely to drop below the specified threshold. In order to compensate for the reduced number of returns, the fill-factor value for objects at longer distances can be calculated by accumulating returns over multiple frames. Allowing points to accumulate over even a short period of time can greatly improve the shape detail in the projected cluster, and increase the fill-factor value appropriately, see Fig. 4. In many instances, false alarms due to random clutter do not persist for more than a frame, therefore their 2D density does not accumulate and their fill-factor value remains low. Although the fill-factor feature helps to eliminate false alarms due to random clutter, it is not able to distinguish humans from other human-sized objects such as barrels and posts. To make this distinction, we are developing a more advanced shape-based technique that involves a 2D comparison of the Ladar returns with a pre-determined ideal human template.

2) Strength-of-Detection

We have developed and tested an effective strength-of-detection (SoD) value to associate with each cluster in each frame of the Ladar data. Initially, we used the fill-factor value as this measure, but this, by itself, does not reflect the increased confidence in the classification that results from repeatedly detecting and labeling a cluster as human over

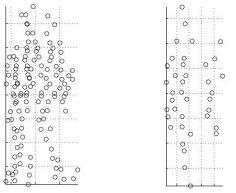

Fig. 3: Projection onto the yz-plane of the Ladar returns from a human at (left) *x* meters compared to (right) *2x* meters.

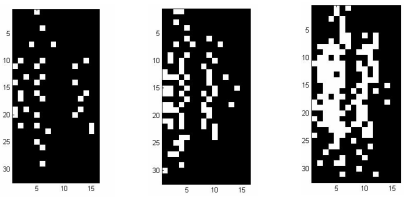

Fig. 4: Improvement in shape information by accumulating Ladar points for clusters at longer range: (left) 1 frame; (center) 2 frames; (right) 5 frames.

several frames of data. Consequently, we implemented an improved SoD measure that takes into account, not only, the fill-factor feature, but whether the object is of the proper size to be human, whether it is moving at a realistic human speed (or is stationary), and whether the cluster persists over time. Confidence is defined in terms of four components: size (c_S), shape (c_f), speed (c_v), and life (c_l),

$$C = c_S * c_f * c_v * c_l , \tag{3}$$

where each component ranges in value from 0 to 1.

The size component, c_S, indicates how closely the height (*h*), width (*w*), and depth (*d*) of each cluster match nominal human dimensions (w_{max}, d_{max}, h_{min}, h_{max}):

$$c_S = \frac{1}{3}(t_W + t_d + t_h) \tag{4}$$

where t_w, and t_d, are defined as

$$t_i = \begin{cases} 1.0, & \text{if } i \le i_{\max} \\ i_{\max} / i, & \text{otherwise} \end{cases}, \qquad i = w \text{ or } d \tag{5}$$

and t_h is

$$t_h = \begin{cases} 1.0 & \text{if } h_{\min} \le h \le h_{\max} \\ h_{\max} / h & \text{if } h > h_{\max} \\ h / h_{\min} & \text{if } h < h_{\min} \end{cases} . \tag{6}$$

We use a sigmoid to define shape in terms of a weighted version of the fill-factor,

$$c_f = \frac{1.0}{1.0 + \exp(-\omega_f - ff * l_h)}, \tag{7}$$

where *ff* is the fill-factor computed from the 2D binned density, l_h is the number of frames the cluster has been human-sized, and ω_f is an experimentally determined constant. Scaling *ff* by l_h increases c_f for clusters that are consistently human-sized.

Since humans are limited in the speed at which they can travel, speed is another feature contributing to our confidence measure. A cluster moving faster than a practical human pace is more likely to be a part of a vehicle than a human, so, as speed increases over the maximum allowable (v_{max}), confidence decreases,

$$c_v = \begin{cases} 1.0, & \text{if } v \le v_{\max} \\ v_{\max} / v, & \text{otherwise} \end{cases}. \quad (8)$$

The final component of C captures the confidence associated with tracking an object and its features over time. The longer a cluster is tracked and is human-sized the more confident we are that we have correctly identified a human. This persistence is characterized by

$$c_l = \frac{1.0}{1.0 + \exp(-\omega_l - l_h * \frac{(l_h - 1)}{l})}, \quad (9)$$

where l is the life of the cluster which is the total number of frames that the cluster has been tracked, l_h is the number of frames the cluster has been human-sized, and ω_l is an experimentally determined constant.

The SoD characterizes the algorithm's confidence that a cluster is human (higher values equal more confidence) and provides a means for doing a thorough receiver operating characteristic (ROC) cuve analysis of algorithm performance.

E. ROC Analysis

We use a baseline set of Ladar data for algorithm analysis that consists of 18 scenarios, each of which is 60 – 90 seconds in duration. In these scenarios, eight humans move along straight line tracks at various orientations to the sensor vehicle. Each particular human always traverses the same basic track. GPS ground truth for each human track, as well as for the sensor vehicle, was recorded at approximately 0.1 second intervals. In addtion to the moving humans, four mannequins were used to represent stationary humans, resulting in a total of twelve possible human targets. Factors that are varied include, the speed of the humans (1.5 or 3.0 meters per second) and the speed of the sensor vehicle (15 or 30 kilometers per hour). As an added challenge, most of the scenarios contain other moving vehicles which occlude the humans from the sensor at certain times.

We define two metrics to characterize the overall performance of our algorithm: the probability of detecting a true track and the number of false tracks generated per second. A track is defined as any set of M points that are labeled with the same ID by the algorithm, whether stationary or moving. As a result, the maximum number of tracks that can be detected for any given scenario is twelve. We typically set the number of points, M, equal to one. Using the SoD defined in (3) thru (9), a threshold, C_T, is defined such that each object with a SoD greater than C_T is classified as human. ROC curves, which plot the probability of detecting true tracks versus the number of false tracks detected per second, are generated for different scenario conditions by varying C_T. Fig. 5 shows the overall performance of our algorithm on the baseline data. Separate

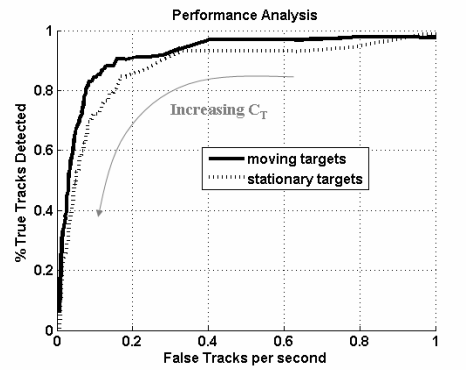

Fig. 5: Average performance results for a set of 18 Ladar files.

curves for the moving and stationary targets reveal the high level of performance that the algorithm achieves in both cases. On average, the algorithm achieves 99% detection with less than one false alaram per second.

III. LWIR ALGORITHM

A. Cluster Extraction and Classification

Our LWIR algorithm generates human regions-of-interest (ROIs) using a two-stage process that first extracts clusters and then classifies the clusters based on a small number of geometrical features. We use a combination of statistical and edge features along with morphological techniques for cluster extraction. Global and local normalized intensity deviation images are computed. Normalized intensity deviation is defined as $n_{ij} = (x_{ij} - m)/\sigma$, where x_{ij} is pixel intensity and m and σ are the mean and standard deviation which have been computed either globally or within a small window around the pixel. For local processing, an integral image implementation has been used so that varying the window size does not impact the computational load of the algorithm.

Edge information is obtained using the gradient. Using empirically determined thresholds, binary images are created from the global and local deviation images, as well as the edge map. Future work is aimed at using statistical processing to automate the choice of the thresholds in each case. Morphological dilation and cleaning are used prior to nearest neighbor clustering.

The second stage of the algorithm computes simple geometrical features for each cluster which are used to retain only those clusters that are human-like in nature. The two features are an axis ratio and an edge ratio. The first is the ratio of the major axis of the cluster to the minor axis of the cluster and the second is the ratio of the number of perimeter pixels in the cluster to the number of edge pixels. We define two thresholds, E_T and A_T, such that all clusters with an axis ratio less than A_T and an edge ratio greater than E_T are classified as human. The features and thresholds have been chosen so as not to eliminate the detection of prone and other non-ideally postured humans.

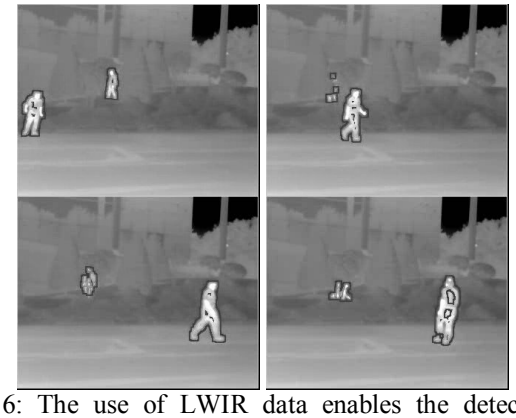

Fig. 6: The use of LWIR data enables the detection of humans in a variety of positions (upright, squatting, visibly occluded).

B. Processing Results

To establish a baseline performance for the algorithm, we first focused on analyzing data sets that are ideal in the sense that the humans tend to be hotter than their surroundings and there is little thermal clutter in the scenes. Fig. 6 shows the results of the algorithm on a series of frames from one of the baseline data sets. Under these ideal conditions, the algorithm provides a high detection rate in conjunction with few false alarms. The algorithm detects upright humans and humans in non-ideal positions (squatting, crouched), as well as humans visibly occluded by vegetation and bushes. In the latter case, the person would not be detectable in EO imagery.

A challenge to our LWIR algorithm performance is the variation in resolution of the data with range. It can actually be more challenging to detect a person at close range than at far range. The lower resolution, at far range, leads to a homogeneous thermal signature for the entire subject, whereas, at close range, the better resolution reveals more variation in a person's thermal signature as a result of things such as clothing and hair, see Fig. 7. At close range, a person often gets separated into smaller features such as the face, arms, hands and legs. In the absence of additional information, it can be challenging to put these pieces back together. However, by incorporating knowledge about the range of the object from the sensor, merging of object pieces can be achieved.

C. Thermal Thresholding Analysis

Like many IR algorithms, our initial algorithm relied on the human having an emissivity greater than that of its surrounding which may be an acceptable assumption at night, and in the early morning, or in cooler climates. However, in general this is not the case, and the algorithm needs to be made robust to various environmental conditions. We are evaluating the ability to thermally threshold the data by leveraging the fact that the sensor is thermally calibrated. However, preliminary analysis has shown that humans generate a wide range of thermal emissions across their body due to variation in clothing material as well as thermal reflections from the surrounding environment. Fig. 8 highlights that a simple windowing

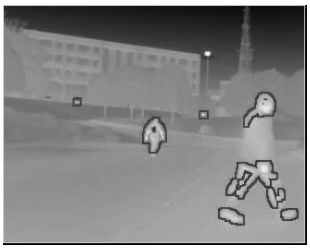

Fig. 7: Variation in thermal resolution with range. At close range, better resolution results in over segmentation, whereas the lower resolution at long range leads to a homogeneous signature across the body.

threshold based on the minimum and maximum intensity of the human eliminates only a small portion of the scene. In this case, all pixels with intensities higher than the maximum human intensity are shown in white and those with intensities less than the minimum human intensity are shown in black. Only the road and sky are eliminated from the scene. In addition, one can see that the legs of the human appear much hotter than the rest of the body due to thermal reflections from the road (road is approx. 120 degrees). We are in the process of developing a more advanced technique which uses statistics of the thermal variation to dynamically threshold the scene and reduce the amount of clutter that remains.

IV. APPROACH TO FUSION

Analysis of our Ladar and LWIR algorithms has revealed the strengths and weaknesses of each approach. Ladar sensors provide a forum for fast and reliable detection and tracking of objects that are well separated in angle and depth, but the data are limited in range and not appropriate for more difficult scenarios (e.g. prone humans). In contrast, LWIR imagery is sensitive to much longer ranges and has the ability to discriminate objects in close proximity as a consequence of emissivity variations. However, due to better resolution at close range, this same emissivity feature leads to over-segmentation of many objects. Robust and reliable human detection and tracking can be achieved by merging these complementary modalities. We have developed a multi-level approach to fusion that combines information from the Ladar

Fig. 8: Thermal emission variation across the human body. Regions hotter than the maximum human intensity are shown in white, cooler in black, and regions which fall between the minimum and maximum are in grayscale with dark gray being the hottest.

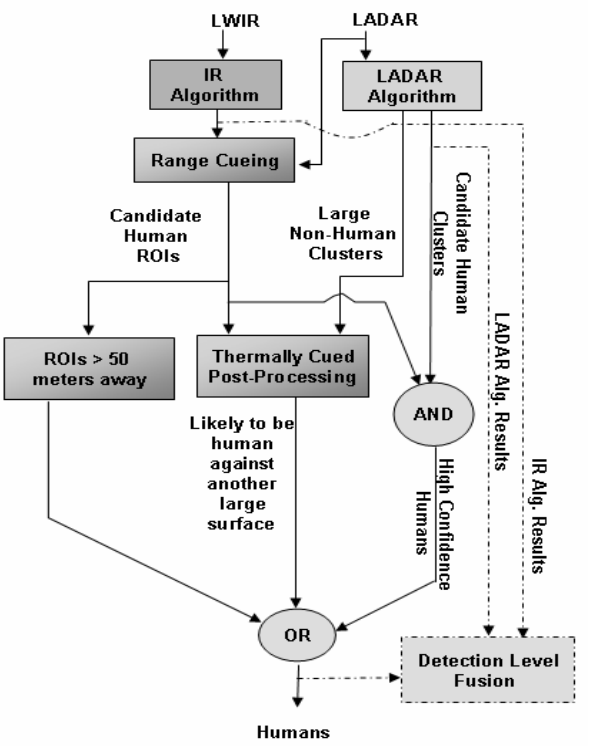

Fig. 9: Approach to fusion includes merging information at both the feature and detection levels.

and LWIR sensors at both the feature and detection levels. Fig. 9 shows an overview of our approach. We are currently addressing the first stage of feature-level fusion in which we incorporate range information with the LWIR imagery.

The most common way of associating depth information with each cluster is through stereo techniques [10], [11]. However, dense stereo methods tend to be computationally intense and may require special processing boards in order to achieve the real-time rates that are required for this task. As a result, we take advantage of registered Ladar and LWIR cameras to use depth information from the Ladar to reduce the computational load on the stereo by eliminating all close range regions in the LWIR imagery from the stereo processing. Stereo techniques provide the long range mapping.

Any approach to fusion is dependent on the involved sensors being accurately registered, both intrinsically and extrinsically. Under the sponsorship of the Robotics Consortium, a method for achieving the desired registration has been developed [13]. Using the results of this newly developed registration process, we will be moving forward with the fusion approach which we have designed.

V. Conclusion

We have developed an algorithm for 3D Ladar data that consists of object detection, data association, classification, and tracking. In depth analysis of the algorithm using a set of ground truthed data, shows that the algorithm provides good performance on humans in more ideal positions that are spatially distinct from other objects. A second algorithm developed for LWIR imagery, complements our Ladar approach by enabling the detection of individuals at much longer range and in more difficult positions. Future efforts will be focused on finalizing the fusion of the two methods.

Acknowledgment

Susan M. Thornton thanks the U.S. AMRMC OHRP and Dr. Sue Hill at ARL for their guidance in ascertaining approval to include human subjects in data collections. In addition, the authors thank NIST for the use of their ground truth equipment and the expertise of the personnel running it. As a result, we have a baseline set of fully ground truthed Ladar, LWIR, and visible wavelength video for algorithm assessment.

References

[1] D.D Morris, B.Colonna, and P. Haley, "Ladar-Based Mover Detection from Moving Vehicles," in *Proceedings of the 25th Army Science Conference,* Nov 27-30, 2006.

[2] L. Navarro-Serment, C. Mertz, and M. Hebert, "Predictive Mover Detection and Tracking in Cluttered Environments," in *Proceedings of the 25th Army Science Conference,* Nov 27-30, 2006.

[3] L. Brown, "View Independent Vehicle/Person Classification," in *ACM International Workshop on Video Surveillance and Sensor Networks 2004,* October 15, 2004, New York, NY.

[4] N. Dalai and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Vol. 1, pp. 886-893, June 20-25, 2005.

[5] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, "Pedestrian Detection using Infrared images and Histograms of Oriented Gradients," in *Intelligent Vehicles Symposium 2006*, June 13-15, 2006, Tokyo, Japan.

[6] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "A Trainable System for People Detection," in *International Journal of Computer Vision,* Vol. 38, No. 1, pp. 15-33, June 2000.

[7] B. Leibe, E. Seemann, and B. Schiele, " Pedestrian Detection in Crowded Scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Vol. 1, pp. 878-885, June 20-25, 2005.

[8] D. M. Gavrila and S. Munder, "Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle," in *International Journal of Computer Vision,* Vol. 73, No. 1, pp. 41-59, June 2007.

[9] J. W. Davis and M. A. Keck, "A Two-Stage Approach to Person Detection in Thermal Imagery," in *Workshop on Applications of Computer Vision, Breckenridge, Co, January 5-7, 2005.*

[10] M. Bertozzi, E. Binelli, A. Broggi, and M. Del Rose, "Stereo Vision-based approaches for Pedestrian Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 3, June 20-26, 2005.*

[11] A. Talukder and L. Matthies, "Real-time Detection of Moving Objects from Moving Vehicles using Dense Stereo and Optical Flow," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems,* Vol. 4, pp. 3718-3725, Sept. 28 – Oct. 2, 2004.

[12] G. Monteiro, C. Premebida, P. Peixoto, and U. Nunes, "Tracking and Classification of Dynamic Obstacles Using Laser Range Finder and Vision," in *IEEE IROS Workshop on Safe Navigation in Open Environments, October 10, 2006, Beijing, China.*

[13] G. W. Sherwin, P. Haley, and M. Hoffelder, "Multi-Sensor Calibration and Registration in Support of Sensor Fusion for Human Detection," AUVSI, June 2008, submitted for publication.

[14] J. Zeng, A. Sayedelah, M. Chouikha, T. Gilmore and P. Frazier, "Infrared Detection of Humans in Stretching Poses Using Heat Flow," submitted to *ICASSP 2008*.