# Ground Truth and Benchmarks for Performance Evaluation

Aya Takeuchi, Michael Shneier, Tsai Hong, Tommy Chang, Chris Scrapper,
Gerry Cheok
National Institute of Standards and Technology
Gaithersburg, MD 20899

## ABSTRACT

Progress in algorithm development and transfer of results to practical applications such as military robotics requires the setup of standard tasks, of standard qualitative and quantitative measurements for performance evaluation and validation. Although the evaluation and validation of algorithms have been discussed for over a decade, the research community still faces a lack of well-defined and standardized methodology. The range of fundamental problems include a lack of quantifiable measures of performance, a lack of data from state-of-the-art sensors in calibrated real-world environments, and a lack of facilities for conducting realistic experiments. In this research, we propose three methods for creating ground truth databases and benchmarks using multiple sensors. The databases and benchmarks will provide researchers with high quality data from suites of sensors operating in complex environments representing real problems of great relevance to the development of autonomous driving systems. At National Institute of Standards and Technology (NIST), we have prototyped a High Mobility Multi-purpose Wheeled Vehicle (HMMWV) system with a suite of sensors including a Riegl ladar, General Dynamics Robotics Systems (GDRS) ladar, stereo Charge Coupled Device (CCD), several color cameras, Global Position System (GPS), Inertial Navigation System (INS), pan/tilt encoders, and odometry[*]. All sensors are calibrated with respect to each other in space and time. This allows a database of features and terrain elevation to be built. Ground truth for each sensor can then be extracted from the database. The main goal of this research is to provide ground truth databases for researchers and engineers to evaluate algorithms for effectiveness, efficiency, reliability, and robustness, thus advancing the development of algorithms.

**Keywords:** Performance evaluation, sensory processing, ladar, ground truth, mobile robots.

## 1. INTRODUCTION

Historically, performance evaluation has not been commonly practiced in the perception community. Periodically, efforts are made to persuade researchers to provide performance evaluations that can be substantiated, but only a few take up this challenge. As a result, performance evaluation is ad hoc in general, and quite frequently completely absent from research papers.

In Europe, a number of formal programs have been developed that address performance evaluation of vision algorithms. Of these, European Computer Vision Network (ECVnet), an association of European vision researchers, had a subcommittee on Benchmarking and Performance Measures[1], although it now appears to be defunct. The German Association for Pattern Recognition (DAGM) established a Working Group on "Quality Evaluation of Pattern Recognition Algorithms", but it, too appears inactive[2]. The International Association for Pattern Recognition has a Technical Committee on Benchmarking & Software, which organizes performance competitions comparing algorithms for particular applications, such as fingerprint identification and document analysis[3]. There have also been a number of workshops on performance characterization and benchmarking of vision systems.

There are also a number of publications that address the issue of how to evaluate the performance of vision algorithms, and a few examples of careful evaluations of particular algorithms or classes of algorithms. Approaches to performance evaluation can be classified into the following general categories, recognizing that more than one approach may be used in an evaluation.

<u>Comparative</u> Here an algorithm may be compared with others that attempt to address the same image-processing task, or its performance may be compared to "ground truth," or perhaps to human performance[4, 5, 6, 7, 8, 9]

[*] Certain commercial equipment, instruments, or materials are identified in this paper in order to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the materials or equipment identified are necessarily best for the purpose.

Analytic The theory behind the algorithm is examined to try to determine the limits to its operation. The computational complexity may be derived, or theoretical optimality may be determined under certain constraints. Frequently, the approach makes use of simplified input data to make the analysis feasible[10, 11, 12, 13].

Performance The way the algorithm actually performs on test data is measured and execution times with different parameters may be reported[14, 15, 16].

Appropriateness to Task The algorithm is shown in the context of a particular application, and the constraints of the task are used to justify the selection of the particular algorithm. The performance of the task as a whole is taken as the evaluation of the algorithm[17, 18].

Other, more informal measures include generality and acceptance. Perhaps the only real performance evaluation measure in common use is longevity. Algorithms that are accepted widely and implemented by many people for different applications can be considered good performers.

A large number of papers report excellent performance of their algorithms, based on small data sets. The success of the Facial Recognition Technology (FERET) program[9] has inspired us to take up the challenge of producing a large database of ground truth for the domain of mobile robotics. In this domain, sensors are mounted on board the moving vehicle, and the algorithms are constrained to run in real time (i.e., fast enough to provide data to control the vehicle). The ground truth that we provide is much more extensive than is typically available, and where human interpretations provide the ground truth, they cover a large number of image sequences because the annotation of the images is performed with computer assistance. We describe three methods for generating ground truth and demonstrate performance evaluations for range and electro-optical sensors using the ground truth database.

We have developed a rigid and reliable methodology for producing three different kinds of large databases of sensor data with ground truth. One method involves collecting ground truth data using a highly accurate ladar sensor mounted on our instrumented HMMWV. The ladar can characterize large areas of terrain and is registered with cameras that provide color information for each ladar point. The position and time at which each sample is collected is recorded with an INS and GPS accurate to a few centimeters. Another set of data was obtained through a high-resolution aerial survey of the grounds of the NIST and surrounding area. The survey includes annotations providing labels for all the features. Lastly, we have developed an interactive method of hand-labeling features in image sequences to efficiently generate a large database of ground truth data.

The data sets are used to evaluate performance of algorithms objectively by comparing the output of the algorithms to the expected result derived from the ground truth. Given a large number of ground truth data sets from different environments, statistical evaluations are possible as well as the robust assessment of performance of algorithms.

The main goal of this work is to make our test data and ground truth available for general use, with the hope that it will lead to rapid and significant development of perception algorithms for autonomous mobility. In order to validate the approach we use the data sets to evaluate our own algorithm development.

## 2. APPROACH

In the following, we assume that all sensors used to collect data and produce ground truth have been calibrated and necessary parameters such as the optical center of a device or the focal plane are known.

We first discuss our method for creating ground truth databases for sequences of color image data. It involves a human user, who annotates the data to supply the ground truth. Manually annotating sensor data with ground truth is costly and time consuming. Instead, we have developed a semi-automatic ground truth application that reduces cost and time by requiring only occasional annotation. The user annotates the first image of a sequence by outlining and naming regions of interest (e.g., highway signs, vehicles). The computer then tracks the annotated regions through successive images, and the user observes how well each region is recognized and outlined by the computer. When the annotations start to diverge from the desired regions, the user intervenes and re-identifies the regions, retaining the same names. When new regions appear that the user wants to track, the same process of stopping the computer, annotating the regions, and restarting the tracking is followed. The annotation application can be used to outline regions with curved or polygonal lines, and several tracking algorithms can be used, depending on the objects in the images. The output of this process consists of the names, shapes and position coordinates of the targets in each image

Figure 1 shows the starting frame of a sequence of color images. It shows road edges that were selected by a user constructing the ground truth. Figure 2 shows the

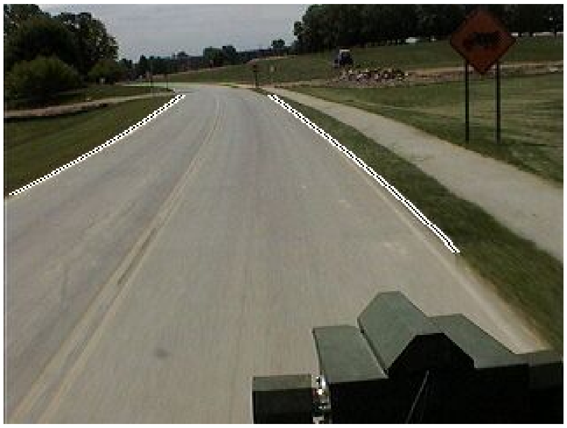

**Figure 1** *The first frame of a sequence. The user has drawn the features to be tracked.*

results of automatic tracking. The tracking to this point is acceptable, and no user interaction is required. Figure 3 shows the situation when the automatic tracking is starting to drift. At this point, the user stops the tracking, resets the annotation, and lets the tracker continue (Figure 4).

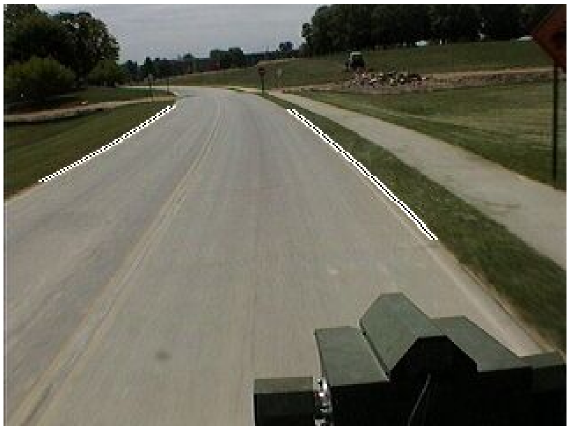

**Figure 2** *The computer tracks the features through a sequence of images*

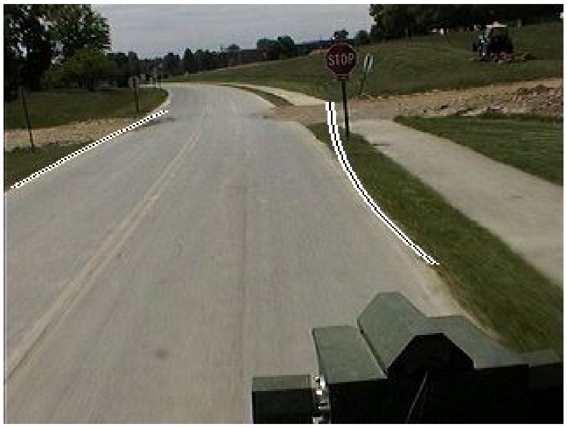

**Figure 3** *In this frame, the automatic tracker has drifted enough to require human intervention.*

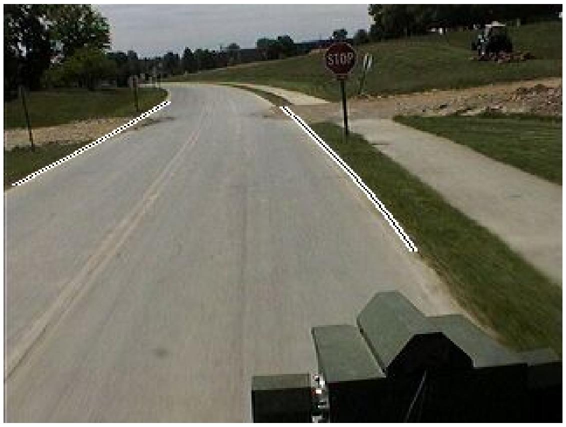

**Figure 4** *The user re-initializes the features and automatic tracking continues.*

The second method provides data for evaluating range sensors. It makes use of a high-resolution ladar (Riegl LMS Z210) to construct a map of a region. The map can then be used for evaluating range sensors that have significantly lower resolution than the Riegl. We use a 5 cm x 5 cm spatial resolution grid to construct the ground truth map, but maps can be constructed at different resolutions (finer or coarser). This method has been used to gather ground truth for off-road terrain such as that shown in Figure 5.

Evaluating other range sensors involves mapping their data into the high-resolution map. The residual of the Riegl data and the other sensor data provides a measure

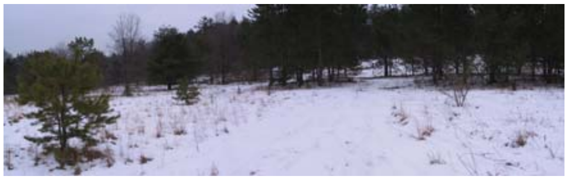

**Figure 5** *An image mosaic used to provide color information for the high-resolution ladar scanner.*

of the performance of the sensor (relative to the Riegl). It is important to note that in order to map data from the sensor under test onto the Riegl data, the positions and orientations of the sensors must be known accurately. The current map resolution of 5 cm x 5 cm corresponds to a spatial tolerance of 5 cm. This method of constructing a map of a region can also measure how much information each successive ladar image adds about the world. The ground truth maps can also be used to evaluate similar maps constructed with stereo algorithms[6]

Figure 6 shows the result of scanning the region in Figure 5 with the Riegl ladar. Figure 8 shows the sub-region scanned with a different ladar (GDRS). In Figure 7 the two scans are overlaid. The white region shows the mismatch due to the lower resolution and coarse range quantization of the GDRS ladar with a small component due to registration error.

The third method  involves constructing a ground truth database of color and range images based on a high-resolution aerial survey combined with data from calibrated ground sensors such as cameras and ladars. In our case, we commissioned a survey of the NIST campus (2.34 * $10^6$m$^2$ or 578 acres) and part of the surrounding urban area. The area includes roads, parking lots, traffic signs, buildings, trees, streams, fences, etc., as well as off-road terrain. All of these features are recorded and entered into a database of features and terrain elevation. Ground truth for each sensor can then be extracted from the database based on position and sensor model.

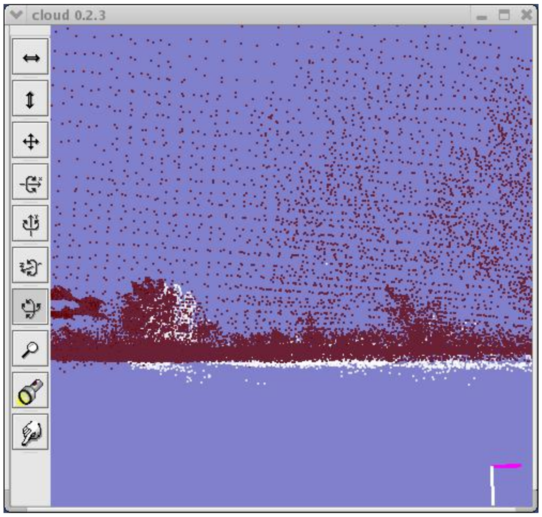

**Figure 7** *The result of overlaying the GDRS ladar data on the Riegl data. The difference in measurement of the scene can clearly be seen.*

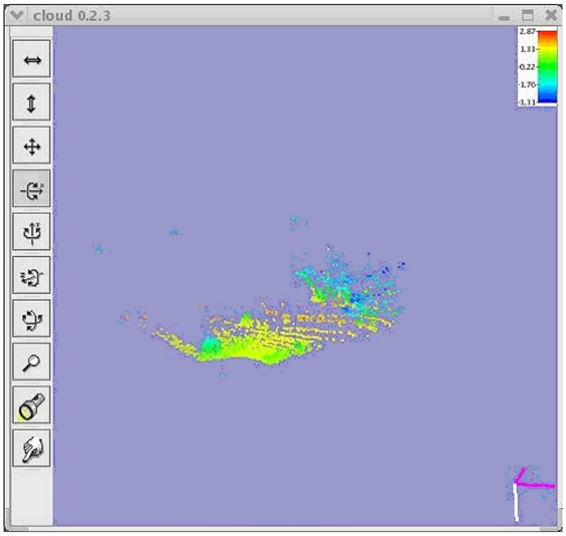

**Figure 8** *The sub-region seen by the GDRS ladar, taken from the same position. Elevation is again represented by color.*

In order to produce the ground truth database from a sensor or set of sensors, each sensor is mounted on the NIST HMMWV or other calibrated vehicle with an accurate position sensor, or a tripod or other stationary mount whose position can be obtained accurately. If the sensors are to be used in real time autonomous driving, the vehicle is driven over the NIST grounds, preferably over the kind of terrain on which the sensors will be used and data is collected with associated time and position stamps[19].

A high resolution INS/GPS navigation sensor coupled with a differential GPS base station and post-processing of the position data enables determining the location of the NIST vehicle with an accuracy of about 4 cm in position and a few thousandths of a degree in orientation. The location of each sensor on the vehicle can also be precisely measured using the techniques described in a companion paper in this proceedings[19]. This enables sensor data to be transformed into the vehicle coordinate system, or into the coordinates of the aerial survey of the NIST campus. The labels of sensed data can then be obtained from the ground truth

With this procedure, a large set of ground truth database can be produced easily. Given a dataset captured in this way, we can borrow the evaluation procedure from FERET program[9] to quantitatively evaluate the performance of sensor-processing algorithms such as segmentation, classification, and recognition algorithms. These algorithms produce labeled regions in an image. The regions, by projection into the a priori data, can be assigned labels from the ground truth. It then becomes a simple matter to determine the percentage of false positive and false negative labels of each algorithm and the correctness of the detected positions and shapes of the objects.

The ground truth data are also an excellent resource for verifying the accuracy of a ladar sensor by taking samples from locations that contain surfaces or objects of known sizes, distances, and orientations. The response of the algorithm is then compared with the ground truth position, which is extracted from the database of prior knowledge based on the known position of the sensor and its field of view. Obviously, all measurements are limited by the accuracy of the a priori data and the accuracy with which the position and orientation of the sensor can be established with respect to the a priori data. A sample-by-sample measurement can be made, giving the range resolution and field of view of the sensor. Alternatively, feature-based measurements can be made, giving the accuracy with which the sensor can capture surfaces of different shapes and slopes. More detailed studies, such as trying to determine which part of the field of view of a single sample (e.g., laser beam) gives rise to the measured response, can also be made, but methods customized to the sensor are more reliable.

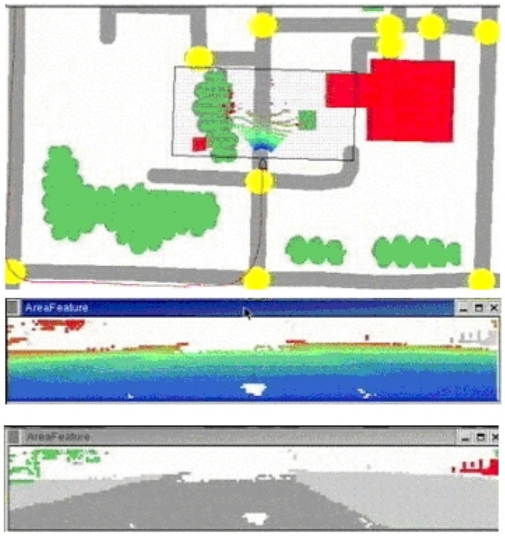

**Figure 9** ***Top*** *A view of the path of the NIST HMMWV traversing the NIST grounds. The features are derived from the ground truth indexed by vehicle position. The rectangle represents the field of view of the GDRS ladar.* ***Middle.*** *A GDRS ladar image taken from the HMMWV.* ***Bottom****. Predicted image from ground truth. Each pixel contains a label predicting the identity of the corresponding location in the ladar image.*

## 3. DISCUSSION AND CONCLUSIONS

Three methods of producing and using ground truth data have been presented and applied to electro-optical and range sensors primarily used for autonomous mobile robots. The methods all rely on ground truth and are dependent on the accuracy with which it is represented and registered with the test samples. By careful measurement of the positions and orientations of the sensors at the time samples are taken, a good match with the ground truth can be established and quantitative measures of performance for sensors and sensory processing algorithms can be made.

We have developed a reliable methodology for establishing a large database of ground truth for evaluating sensors and sensor-processing algorithms. The database is available to the public with the hope that researchers and engineers will use it to verify and evaluate sensors and algorithms for effectiveness, efficiency, reliability, and robustness. This will enable algorithms to be developed using realistically difficult sensory data, make it possible to compare algorithms quantitatively by running them on the same data, and speed technology transfer by providing industry with metrics for comparing algorithm performance. It will also help with sensor development by highlighting areas of strength and weakness of current sensors.

### Acknowledgements

We are grateful for the support of the Army Research Laboratory and the DARPA MARS project.

## REFERENCES

1. Courtney, P., Benchmarking and Performance Evaluation , http://www-prima.inrialpes.fr/ECVNet/benchmarking.html, Mar.,1998.

2. Faber, A., Quality Characteristics of Pattern Recognition Algorithms,

http://www.dagm.de/DAGM/ag/wg.html, May,1998.

3. Lucas, S., IAPR TC-5 Benchmarking and Software, http://algoval.essex.ac.uk/tc5/Introduction.html, 2003.

4. Bowyer K., Kranenburg, C., and Dougherty, S., "Edge Detector Evaluation Using Empirical ROC Curves," *Proceedings of the IEEE COmputer Society Conference on Computer Vision and Pattern Recognition* 354-359, IEEE, Los Alamitos, CA.

5. Nguyen, T. B. and Zhou, D.,"Contextual and Non-Contextual Performance Evaluation of Edge Detectors" *Pattern Recognition Letters* **21**, 805-816, 2000.

6. Matthies, L., Litwin, T., Owens, K., Murphy, K, Coombs, D., Gilsinn, J., Hong, T., Legowik, S., Nashman, M., and Yoshimi, B., "Performance Evaluation of UGV Obstacle Detection with CCD/FLIR Stereo Vision and LADAR," *IEEE Workshop on Perception for Mobile Agents*, Santa Clara, CA.

7. Shufelt, J. A.,"Performance Evaluation and Analysis of Monocular Building Extraction From Aerial Imagery" *IEEE Transaction on Pattern Analysis and Machine Intelligence* **21**, 311-326, 1999.

8. Wiedemannm C., Heipke, C., Mayer, M., and Jamet, O., "Empirical Evaluation of Automatically Extracted Road Axes." *Empirical Evaluation Techniques in Computer Vision*, 172-187.

9. Phillips, P. J., Moon, H., Rizvu, S. A., and Rauss, P.,"The FERET Evaluation Methodology for Face-Recognition Algorithms" *IEEE Transaction on Pattern Analysis and Machine Intelligence* **22**, 2000.

10. Cho, K., Meer, P, and Cabrera, J,"Performance Assessment Through Bootstrap" *IEEE Transaction on Pattern Analysis and Machine Intelligence* **19**, 1185-1198, 1997.

11. Courtney, P., Thacker, N., and Clark, A. F.,"Algorithmic Modelling for Performance Evaluation" *Machine Vision and Applications* **9**, 219-228, 1997.

12. Kiryati, N., Kälviäinen, H., and Alaoutinen, S,"Randomized or Probabilistic Hough Transform: Unified Performance Evaluation" *Pattern Recognition Letters* **21**, 1157-1164, 2000.

13. Haralick, R., "Propagating Covariance In Computer Vision," *Proceedings of the ECCV Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, England.

14. Pissaloux, E. E.,"Toward an image segmentation benchmark for evaluation of vision systems" *Journal of Electronic Imaging* **10**, 203-212, 2001.

15. Min, J., Powell, M. W., and Bowyer K., "Automated performance evaluation of range image segmentation," *Fifth IEEE Workshop on Applications of Computer Vision*163-168, IEEE, Palm Springs, CA.

16. Coutre, S. C, Evens, M. W., and Armato II, S. G., "Performance Evaluation of Image Registration," *Proceedings of the 22nd Annual EMBS International Conference*3140-3143, IEEE, Chicago, IL.

17. Shin, M. C., Goldgof, D., and Bowyer, K. W., "Objective Comparison Methodology of Edge Detection Algorithms Using a Structure From Motion Task," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*IEEE, Santa Barbara, CA, 1998.

18. Moon, H., Chellappa, R., and Rosenfeld, A.,"Performance Analysis of a simple Vehicle Detection Algorithm" *Image and Vision Computing* **20**, 1-13,2002.

19. Shneier, M. O, Chang, T., Hong, T., and Cheok, G. Scott H., "A Repository Of Sensor Data For Autonomous Driving Research," *SPIE Aerosense Symposium.*