

Combining Laser Range, Color, and Texture Cues for Autonomous Road Following

Christopher Rasmussen*

National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

We describe results on combining depth information from a laser range-finder and color and texture image cues to segment ill-structured dirt, gravel, and asphalt roads as input to an autonomous road following system. A large number of registered laser and camera images were captured at frame-rate on a variety of rural roads, allowing laser features such as 3-D height and smoothness to be correlated with image features such as color histograms and Gabor filter responses. A small set of road models was generated by training separate neural networks on labeled feature vectors clustered by road "type." By first classifying the type of a novel road image, an appropriate second-stage classifier was selected to segment individual pixels, achieving a high degree of accuracy on arbitrary images from the dataset. Segmented images combined with laser range information and the vehicle's inertial navigation data were used to construct 3-D maps suitable for path planning.

1 Introduction

An autonomous vehicle navigating on- and off-road (e.g., military reconnaissance) must be aware of different kinds of terrain in order to make prudent steering decisions. To maximize safety and speed, it may be desirable to use any roads in an area of operation for as much of a point-to-point path as possible. This special case of general terrain traversal, *road following*, requires an ability to discriminate between the road and surrounding areas and is a well-studied visual task. Much work has been done on driving along highways and other paved or well-maintained roads [1, 2, 3], but marginal rural and backcountry roads are less amenable to standard techniques for a variety of reasons. There may be no lane lines or markings; the road/non-road border is often spatially fuzzy and has low intensity contrast; the overall road shape may not follow smooth curves and the support surface may be highly non-planar; and the appearance of the road

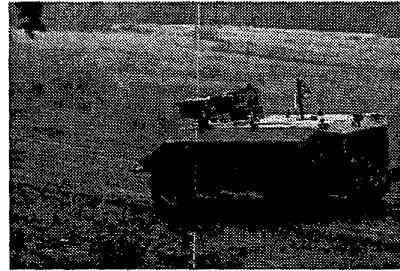


Figure 1: Experimental Unmanned Vehicle (XUV) driving in area where data was collected for this paper.

itself can change drastically: mud, clay, sand, gravel, and asphalt may all be encountered.

Algorithms that attempt to delineate the road via region-based segmentation have been fairly successful. Color [4, 5] and texture [6] are two characteristics that have been used to differentiate the road from bordering vegetation or dirt. Some work has also been done on using 3-D information to constrain segmentation: e.g., [7] applied structure-from-motion techniques to automatically detected and tracked features in order to follow a dirt road in the midst of dense trees. Visual and structural modalities are clearly complementary: vision alone may be inadequate or unreliable in the presence of strong shadows, glare, or poor weather, while road boundaries do not necessarily coincide with 3-D structures—the height border between a dirt road and short grass, for example, is undetectable by most current methods and sensors.

Classification offers a straightforward way to combine these two sources of information. In this paper, we report work on road segmentation using a camera and a laser range-finder mounted on an autonomous four wheel-drive vehicle, the Experimental Unmanned Vehicle (XUV) (shown in Figure 1), which is part of the Army Demo III project [8]. By framing the problem as one of learning by labeled examples whether small image patches (registered with laser range information) belong to the road or background, we can easily integrate disparate features such as 3-D height

*E-mail: crasmuss@nist.gov. This work was performed while the author held a National Research Council Research Associateship Award at NIST.

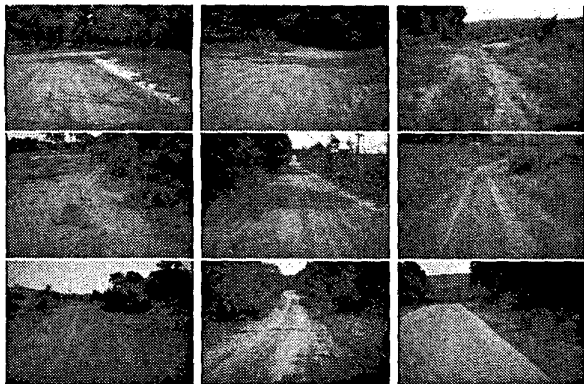


Figure 2: Sample road images

and smoothness with image qualities like color and texturedness. We have found that fusing these modalities yields better performance than any one method over a wide variety of individual road images. Clearly, though, it is infeasible to learn a separate model for every image. Learning a single model for the entire image corpus is a simple solution, but it reduces classification accuracy because of the variety of road and background types that must be handled. Therefore, we propose a method to automatically learn and apply a small number of different road appearance models which boosts performance considerably.

In the next three sections we will briefly describe the background behind our approach, then detail our experimental procedures and training and testing data, and finally present results.

2 Road segmentation

We frame road segmentation as a classification problem in which we wish to identify small patches over the field of view as either road or non-road on the basis of a number of properties, or *features*, that we compute from them. These features are non-geometric: image location is not considered for segmentation, only local image properties. Patches are manually labeled for a representative set of images (Figure 2 shows some examples from our data), and a neural network [9] is trained to learn a decision boundary in feature space. This model is used to classify pixels in novel images, from which we can either (1) derive road shape parameters directly by recursively estimating curvature, width, etc. from the edges of the road region and control steering accordingly (analogous to [3]); or (2) use the laser information to backproject road and non-road regions into a 3-D map (see Section 4 for an example) suitable for a more general path planner [10].

We have two sensors available—a laser range-finder which gives dense depth values and a video camera—with differing fields of view and capture rates. By registering the images obtained from each sensor both spatially and temporally (our procedure is explained in the next section), we can formulate an *image pair* that contains correlated information from both. We have chosen four basic kinds of features to distinguish road patches from plants, rocks, tree, grass, and other off-road zones—two from the laser half of the pair and two from the image half. They are:

Height Vertical distance of laser point from the vehicle support surface.¹ This should allow bushes and trees to be eliminated regardless of visual appearance.

Smoothness The height variance in the neighborhood of a laser point. Roads should be locally flat, while tall grass and loose rocks are bumpier.

Color A color histogram [11] is computed over each image patch. Roads are expected to be more-or-less consistent in their mix of colors—generally brown or gray—while the background is expected to exhibit more green and blue colors to allow discrimination.

Texture Gabor filters [12] are computed over each image patch to characterize the magnitude and dominant direction of texturedness at different scales. The road should be more homogeneous or anisotropic (e.g., tracks, ruts) than bordering plants.

3 Methods

Real-time video, laser range data, and inertial navigation information were recorded on the XUV as it was tele-operated along a variety of dirt and asphalt roads at Fort Indiantown Gap, PA in July, 2001. Data spanning approximately 73 min of late-morning driving at 8-24 km/h was captured in 14 distinct sequences totaling 131,471 video frames.

The analog output of the camera, a Sony DXC-390,² was converted to DV before capture and then subsampled, resulting in a final resolution of 360×240 for image processing. The laser range-finder, a Schwartz SEO LADAR, acquires a 180×32 array of range values at ≈ 20 Hz covering a field of view of 90° horizontally and 15° vertically.

For training, 120 video frames were randomly chosen and the most-nearly synchronous laser range image was paired with each. Of these, nine image pairs were eliminated due to missing data in the laser im-

¹Throughout this paper, +Z is forward with respect to vehicle direction, +X is right, and +Y is up. The height h and tilt angle θ of the sensors are known and accounted for.

²Certain commercial materials and equipment are identified in this paper to specify experimental procedures adequately. Such identification does not imply endorsement by NIST.

age (a hardware artifact) and four because the vehicle was not on a road. This left 107 image pairs for training and testing. Road regions (some roads had two tracks separated by grass) were manually marked in each camera image with polygons.

3.1 Features

Feature vectors were computed for each image at 10-pixel intervals vertically and horizontally, with roughly a 20-pixel margin to ensure that filter kernels remained entirely within the image. This resulted in 640 feature vectors per image. Centered on each feature location, three different sizes of subimage were examined for feature computation: 7×7 , 15×15 , and 31×31 . A total of fourteen *feature sets*, or segments of the full feature vector, were used for learning. These consisted of:

Six color feature sets Two kinds of color features were computed over the above three scales: a standard 4-bins-per-RGB-channel joint color histogram (4^3 total bins), and an “independent” color histogram consisting of 8 bins per channel (8×3 total bins).

Two texture feature sets Texture features consisted of the odd- and even-phase responses of a bank of Gabor filters histogrammed over the 7×7 and 15×15 scales (8 bins per phase with limits defined by the max and min filter response on each particular image). For each phase, the Gabor filter bank consisted of three wavelengths (2, 4, and 8—resulting in kernel sizes of 6×6 , 12×12 , and 25×25 , respectively) and eight equally-spaced orientations.

Six laser feature sets As Figure 3 shows, not every image location has laser information associated with it. Only those feature vectors with adequate laser information (> 1 point projecting into its subimage) were included in training with any feature subset that was not exclusively image-based. For eligible locations, the mean and covariance were computed of the X, Y, Z coordinates of the n laser points projecting to the local 15×15 or 31×31 image neighborhood. As features we used the mean Y value, the variance of Y , and the Y mean and variance over the two scales. The Y mean allows discrimination based on height relative to the base of the vehicle’s tires, while the Y variance was included as a simple measure of smoothness.

3.2 Calibration and classification

The camera’s internal parameters were calibrated using J. Bouguet’s Matlab toolbox [13]. The external orientation between the camera and LADAR was obtained by correlating corresponding points imaged by each device over a number of scenes and then computing a least-squares fit to the transformation according to the procedure described in [14]. A generic model

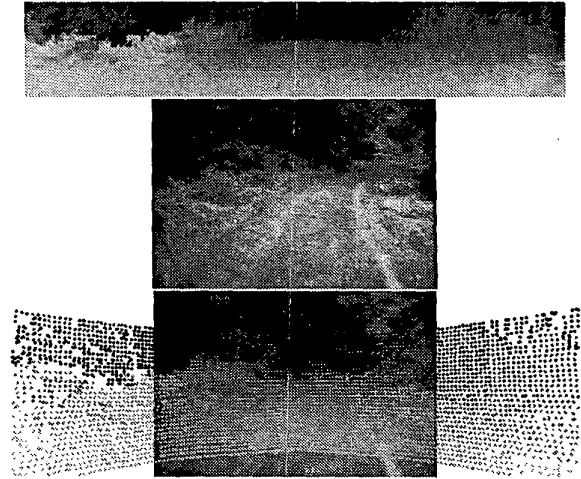


Figure 3: Sample laser-camera registration. Laser pixel distance is proportional to darkness.

was used for the internal calibration parameters of the Schwartz LADAR though they are known to vary fairly significantly from device to device, limiting the accuracy of the camera-laser registration. Rectangular areas of erroneous pixel depths (such as the anomalous stripe in the road in Figure 4(a)) were occasionally introduced by the Schwartz device’s method of acquiring each scene as a series of smaller image facets.

The Matlab Neural Network Toolbox [15] was used to train the neural networks in this paper. Each neural network had one hidden layer consisting of 20 hidden units; weights were updated using conjugate-gradient back-propagation with the “tansig” activation function. During training, the classification accuracy of a particular neural network was estimated using cross-validation, where $\frac{3}{4}$ of any given data set was used as a training *fold* and the remaining $\frac{1}{4}$ for testing, rotating the testing fraction four times. The quoted accuracy is the median of the four testing accuracies.

4 Results

We experimented with a number of different training regimes to assess the utility of the various modalities (laser, color, and texture) both independently and in combination, on individual images and on the sample corpus as a whole.

4.1 One model per image

A separate neural network was trained on each of the 107 random camera-laser pairs $\{I_i\}$ for each of the feature sets described in the previous section. Taking the mean accuracy of each feature subset over all images, the best performers by modality were the 31×31

Features	S	Min	Std	DD	DS	SD	$k = 4$
C	97.0	81.3	3.2	93.7	93.6	75.4	94.8
T	88.6	78.4	3.9	77.8	78.8	52.3	81.3
L	84.8	70.1	5.0	78.1	78.1	69.6	—
C + T	97.3	75.0	2.7	94.7	95.5	62.6	96.1
C + L	96.1	88.0	2.0	89.5	90.2	71.3	91.6
T + L	91.2	81.0	3.7	81.3	81.5	54.2	84.1
C+T+L	96.6	91.2	1.8	91.0	92.8	59.6	93.3

Table 1: Mean feature subset performance for various training and testing regimes. Features: C=color, T=texture, L=laser. Data sets: S=107 individual images; D=25% all-image digest (1st letter=training, 2nd=testing).

independent color histogram, the 15×15 Gabor histogram, and the 31×31 laser Y mean and variance. The percentage mean accuracies over all images for these best individual performers, as well as for feature sets comprising combinations of them (color and texture, texture and laser, etc.) trained in the same way are shown in the S column of Table 1.

Color was clearly the most informative of the modalities, though texture and laser alone did fairly well³. Combining texture and laser features with color did not appreciably change the mean accuracy, but it increased consistency of performance. The standard deviation of the accuracy Std was cut almost in half going from color alone to color, texture, and laser together (C+T+L), and the minimum accuracy Min (i.e., on the image eliciting the worst performance for that feature set) went up nearly 10%. This pattern was repeated for the other modalities, indicating that adding features often resolved scene ambiguities.

For example, each row of Figure 4 shows the most difficult images to classify using laser alone and texture alone. The left column shows the segmentation obtained by the best-performing neural network of the training folds for that individual modality. The right column shows the results of segmenting the same image with the C+T+L classifier's best training fold neural network. The laser classifier's defect in Figure 4(a) is most obvious: the asphalt road and grassy strip to the right are in the same plane and both quite smooth, which is why the segmentation erroneously extends to the treeline on the right. The color and texture discontinuity between the two regions is much clearer in (b).

³As a baseline for performance assessment, the mean proportion of feature vectors labeled "road" over all 107 images was 47.7%. Considering only those feature vectors containing adequate laser information (for the 31×31 subimage size), this fraction was 55.7%.

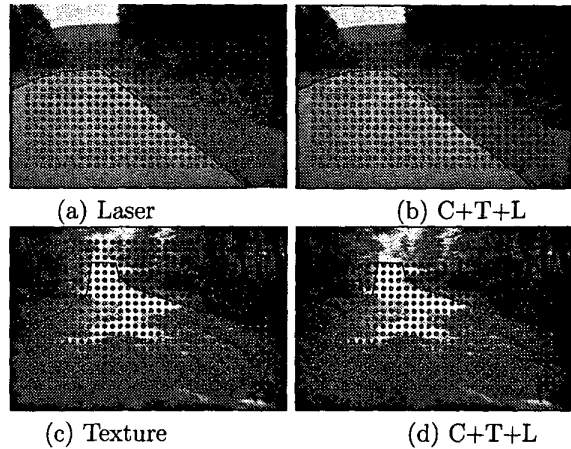


Figure 4: Segmentation of hardest road images for independent modalities vs. joint classifier

The texture classifier presumably has trouble with its image in (c) because of the similar patterns of light and shadow in the trees and on the road; adding color and laser information nearly eliminates these problems.

4.2 One model for all images

To test learning a single road model for the entire corpus as well as the generality of the individual image models, a *digest* D was created from the set of 107 images by randomly selecting 25% of each image's feature vectors and concatenating them. Of D's 17,120 feature vectors, 8,168 or 47.7% were labeled as "road."

Training was performed on D for the seven feature sets from Table 1 exactly as if it were a larger version of an image I_i . Results are shown in the DD column of the table. The power of the digest to faithfully represent the images themselves can be seen in the similarity of the accuracies obtained by training and testing on the digest alone (DD) to training on the digest and computing the mean accuracy over all of the individual images (DS). Performance with a single model for the entire digest declines somewhat across all of the feature sets from the mean accuracy of separate models for every image (S). This effect is most pronounced for texture, indicating that on-road and off-road textures are more similar for the entire image corpus than, say, on-road and off-road colors.

The poor generality of the single-image models learned in the previous subsection is demonstrated by testing them on D; the mean performance over the 107 images is given in column SD of the table. Accuracy drops dramatically because of the presentation of road and background types not seen in the single image training.

As an example of the utility of the laser information beyond segmentation, a road map constructed from one manually-driven sequence over roughly 300 meters (1825 frames) is shown in Figure 5. As the vehicle traveled from the lower-right to the upper-left corner of the map, the image was segmented at 10 frame intervals using the single-model, color-only classifier C. The labels of feature locations with associated laser-derived depths were projected into a 1-meter resolution grid square world map using position information from the vehicle's inertial navigation system. Neglecting height for clarity, the map shows the degree of roadness/non-roadness of each grid square along a green/red continuum, with color saturation indicating confidence (proportional to the number of labels projected to the square, up to 5). White grid squares were not mapped.

Overall, the road is mapped quite clearly despite shadows and changes in road composition. Three difficult views along the route at map positions *a*, *b*, and *c* (blue dot=position, purple dot=viewing direction) are shown in Figures 5(a)-(c). The left road edge is not as sharp as the right at position *a* because the road dirt extends into the trees. Road is found in a large forward area at position *b* because the vehicle is at an intersection before turning right. Finally, the transverse road boundary is easily found on the opposite side of the T-intersection at position *c*.

4.3 One model per road type

The lesser performance of a single neural network trained on a digest of all of the images versus that of individual networks for each image is presumably due in large part to the greater overlap of road and non-road feature vector distributions in the former method's training set. Partitioning a digest \mathbf{D} into pieces $\mathbf{d}_1, \mathbf{d}_2, \dots$ such that the road and non-road feature vector distributions are more widely separated within each \mathbf{d}_i than in \mathbf{D} , then training on each \mathbf{d}_i , would likely reduce the difficulty of the classification problem. Observing that the within-image contrast between road and non-road was strong across the sample images, we made the following important assumption: that similar road types are correlated with similar background types in each image. This implies that clustering road types is roughly equivalent to clustering background types, and that all of the background types within such a cluster would on average be more dissimilar to the road types in the cluster than those of the digest as a whole.

We tested this hypothesis by using *k*-means clustering [16] to group the 107 sample images for the best color feature set C, the best texture feature set

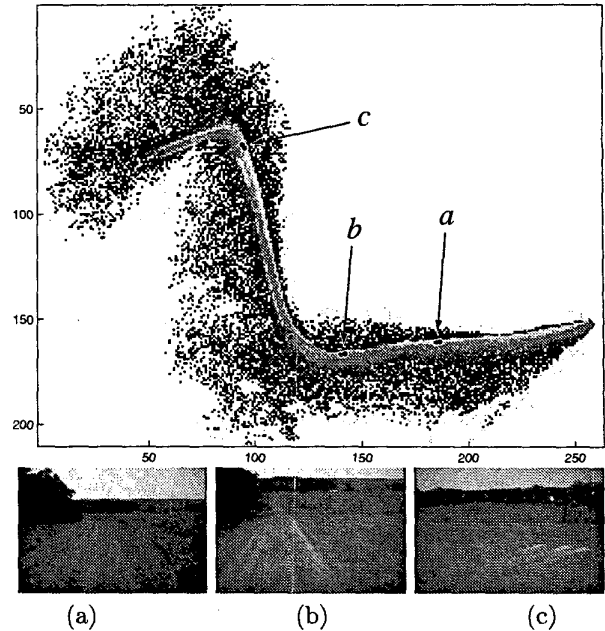


Figure 5: Road map constructed with single-model feature set C classifier. Units are meters.

T, and the best color and texture feature set $\mathbf{C} + \mathbf{T}^4$. Roads were not clustered with laser feature information because the major variation in road types for this data is visual: dirt, gravel, and asphalt have marked differences in color and degree of texturedness, but all roads were approximately smooth and at the same height relative to the vehicle.

Ideally, every road-labeled feature vector in an image would define a "road signature" and thus the space in which clustering is done, but this fails because (a) the number of feature dimensions would exceed the number of sample images, and (b) after training is done and the system is in operation, feature vectors will not be labeled (that being the point of segmentation). First, to reduce the dimensionality principal component analysis [16] was performed on the road-labeled digest feature vectors $\mathbf{R} \subset \mathbf{D}$ to obtain a transformation that orthogonalized feature space and removed those principal components that contributed less than $N\%$ of the variation. A fairly large N was chosen because of the small number of samples (e.g., $N = 15\%$ for C, compressing 24 features down to 2; $N = 4\%$ for T, reducing 384 features to 3; and $N = 3\%$ for $\mathbf{C} + \mathbf{T}$, taking 408 features to 5). Second, a small

⁴The algorithm was run 50 times with random seeds for each $k = 2, 3, 4, 5$ and feature set; the result exhibiting the lowest within-cluster scatter to between-cluster scatter ratio was used.

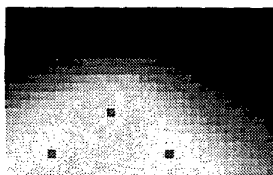


Figure 6: Probability of a feature location being labeled road over sample images, with "road signature" locations overlaid.

subset of feature vector locations was chosen to represent the road signature of each image, as shown by the points in Figure 6, with the goals of (a) maximizing the *a priori* probability of them being labeled road based on the sample images, and (b) an even distribution to capture spatial variation of feature values across the road region.

After clustering for each k , \mathbf{D} was divided into pieces $\mathbf{d}_1, \dots, \mathbf{d}_k$ according to which image each block of 160 feature vectors was taken from, and a separate neural network was trained on each \mathbf{d}_i . For every cluster i , the associated best neural network (i.e., from the training fold with the highest accuracy) was then tested on all of the sample images in that cluster. A consistent performance increase of up to several percentage points over the single-model classifiers in the DS column of Table 1 was obtained across all of the feature sets and values of k , with $k = 4$ (performance shown in the last column of Table 1) yielding the greatest average improvement. The quality of clustering would likely be better with more sample images.

5 Conclusion

We presented a road segmentation system that integrates information from a registered laser range-finder and camera. Road height, smoothness, color, and texture were combined to yield higher performance than individual cues could achieve. By clustering the roads into a few different types and training a neural network for each, accuracy on the entire image corpus was improved over a simple single-model approach while still retaining good generality. Laser range information was invaluable both as a feature for segmentation and for fusing labeled images into a 3-D map, though better laser-camera registration would likely have produced higher classification performance.

The segmentation procedure described here assumes that the vehicle is on a road and is traveling along it. For vehicles which may operate off-road, road detection is a necessary precursor to road following. Using visual and laser feature sets similar to those exploited

here, an additional classifier could be trained to recognize scenes containing roads in order to turn on or off the road segmentation module. Our data set contains GPS position information for the vehicle; combined with an *a priori* map of roads in the vicinity this would provide a strong additional cue for training a road detection classifier.

For maximum generality, the data set used for training needs to be augmented to capture the visual and structural effects of temporal variations such as time of day, weather, and season. Different road models could be learned for these conditions; fewer such models might suffice if parametrized by continuous variables such as sky brightness or sun angle.

References

- [1] E. Dickmanns, "Vehicles capable of dynamic vision," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1997, pp. 1577-1592.
- [2] D. Pomerleau, "RALPH: Rapidly adapting lateral position handler," in *Proc. IEEE Intelligent Vehicles Symp.*, 1995, pp. 506-511.
- [3] C. Taylor, J. Malik, and J. Weber, "A real-time approach to stereopsis and lane-finding," in *Proc. IEEE Intelligent Vehicles Symp.*, 1996.
- [4] J. Crisman and C. Thorpe, "UNSCARF, a color vision system for the detection of unstructured roads," in *Proc. Int. Conf. Robotics & Automation*, 1991, pp. 2496-2501.
- [5] J. Fernandez and A. Casals, "Autonomous navigation in ill-structured outdoor environments," in *Proc. Int. Conf. Intelligent Robots & Systems*, 1997.
- [6] J. Zhang and H. Nagel, "Texture-based segmentation of road images," in *Proc. IEEE Intelligent Vehicles Symp.*, 1994.
- [7] S. Smith, "Integrated real-time motion segmentation and 3D interpretation," in *Proc. Int. Conf. Pattern Recognition*, 1996, pp. 49-55.
- [8] C. Shoemaker and J. Bornstein, "The Demo III UGV program: A testbed for autonomous navigation research," in *Proc. IEEE Int. Symp. Intelligent Control*, 1998, pp. 644-651.
- [9] B. Ripley, *Pattern Recognition & Neural Networks*, Cambridge University Press, 1996.
- [10] D. Coombs, K. Murphy, A. Lacaze, and S. Legowik, "Driving autonomously offroad up to 35 km/h," in *Proc. IEEE Intelligent Vehicles Symp.*, 2000.
- [11] M. Swain and D. Ballard, "Color indexing," *Int. J. Computer Vision*, vol. 7, no. 1, pp. 11-32, 1991.
- [12] T. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 18, no. 10, pp. 959-971, 1996.
- [13] J. Bouguet, "Camera Calibration Toolbox for Matlab," Available at www.vision.caltech.edu/bouguetj/calib.doc. Accessed May 11, 2001.
- [14] M. Elstrom, P. Smith, and M. Abidi, "Stereo-based registration of LADAR and color imagery," in *SPIE Conf. Intelligent Robots & Computer Vision*, 1998, pp. 343-354.
- [15] H. Demuth and M. Beale, "Matlab Neural Network Toolbox User's Guide, v. 4.0," The MathWorks Inc., 2000.
- [16] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed., John Wiley and Sons, 2001.