In **Proc. Conf. on Computer Vision & Pattern Recognition Technical Sketches, Kauai, HI, December, 2001**

Laser Range-, Color-, and Texture-based Classifiers for Segmenting Marginal Roads

Christopher Rasmussen*

National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

We describe preliminary results on combining depth information from a laser range-finder and color and texture image cues to train classifiers 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 color histograms and Gabor filter responses. A small set of road models were 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.

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 minimize terrain-based dangers and maximize speed, it is often desirable to use any roads present 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 [2, 3, 1], 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 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 [5] and texture 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: for example, [4] applied structure-from-motion techniques to automatically detected and tracked features in order to steer a vehicle along 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 ongoing work on road segmentation using a camera and a laser range-finder mounted on an autonomous four wheel-drive vehicle. 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 and smoothness with image qualities like color and texturedness. We contend that fusing these modalities will yield better performance than any single method. Because of the variety of road types that must be handled, we also propose a method to automatically learn different models for disparate characteristics.

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

Road segmentation can be framed 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 patches are manually labeled for a representative set of images (Figure 1 shows some examples from our data), and a neural

*This work was performed while the author held a National Research Council Research Associateship Award at NIST

Figure 1: Sample road images

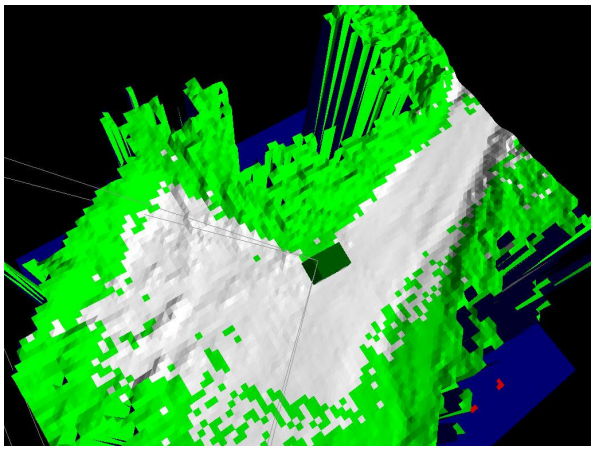

Figure 2: 3-D map with road painted white

network [8] is trained to learn a decision boundary in feature space. This model can be 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 [1]); or (2) use the laser information to backproject road and non-road regions into a 3-D map suitable for a more general path planner, a method we are currently using that is shown in Figure 2.

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 How far a laser point is vertically from the vehicle support surface [1]. This should allow bushes and trees to be eliminated regardless of their 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 [7] is computed over each image patch. Roads are expected to be generally brown or gray, while the background is more green or blue if sky.

Texture Gabor filters [6] are computed over each image patch to characterize the magnitude and dominant direction of texturedness at different scales.

3 Implementation

Real-time video, laser range data, and inertial navigation information were recorded from a robotic vehicle tele-operated on a variety of dirt and asphalt roads at Fort Indiantown Gap, PA in July, 2001. Approximately 73 min of late-morning driving at 8-24 km/h were 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 image (a hardware artifact) and four because the vehicle was not on a road. This left 107 image pairs for training and testing. One contiguous road region was manually marked in each camera image with a single polygon

[1] A vehicle-centric coordinate system is chosen so that $+Z$ is forward with respect to the direction the vehicle is pointing, $+X$ is right, and $+Y$ is up. The height h and tilt angle θ of the camera/laser are known.

[2] Certain commercial equipment, instruments, or materials are identified in this paper to specify experimental procedures adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor that the materials or equipment identified are necessarily the best available for the purpose.

(some "two-track" roads with grass growing down the middle necessitated somewhat contorted boundaries to exclude these areas).

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.

Two kinds of color features were computed over these 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).

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.

Laser features were obtained for only a subset of the total feature locations in an image. For the two largest scales, 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 ($n > 1$).

The camera's internal parameters were calibrated using Bouguet's Matlab toolbox [10]. 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 [11].

The Matlab Neural Network Toolbox [12] 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 for training 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

One model per image From the 107 random camera-laser pairs, one representative of each sequence with the lowest frame number (i.e., earliest in the sequence) was chosen. By chance, every sequence had a representative. Some of the camera images from these 14 pairs $\mathbf{I}_1, \ldots, \mathbf{I}_{14}$ are shown in Figure 1.

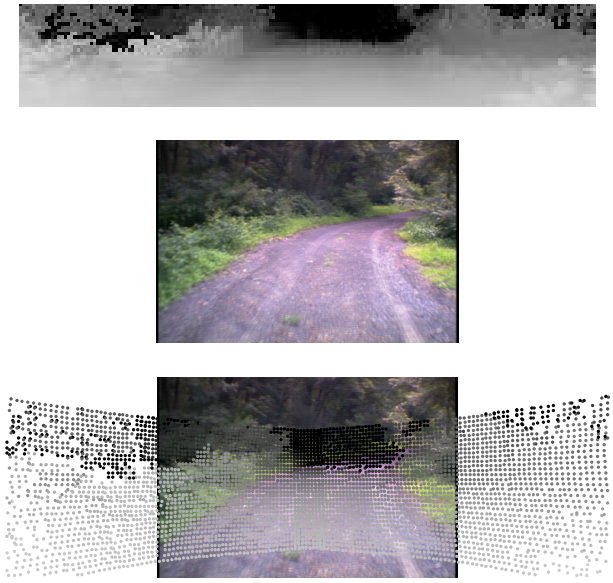

Figure 3: Laser-camera registration

Neural networks were trained on each $\mathbf{I}_i$ using different feature subsets in order to assess the usefulness of color, texture, and laser cues for road classification. In the image domain, these feature subsets included the independent and joint color histograms described in the previous section over all three subimage sizes and the Gabor response histogram over two sizes.

For the laser the mean Y value, the variance of Y, and the Y mean and variance together were examined over two subimage sizes. 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. 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.

Altogether, eight image feature subsets and six laser subsets were tested initially. Taking the median accuracy of each feature subset over all 14 images, the best performers by category were the 15×15 independent color histogram with 97.3%, the 15×15 Gabor histogram with 88.6%, and the 31×31 laser Y mean and variance at 84.6%[3]. All combinations of feature sets

[3]Using principal component analysis to transform the feature space before learning improved performance slightly for color features and decreased it for texture features. The cost

Features	SS%	SA%	AA%	AS%
C + T	97.8	62.1	92.7	94.0
C + L	96.0	74.6	88.4	89.4
T + L	92.2	55.6	78.8	81.6
C+T+L	95.8	60.2	91.4	91.9

Table 1: Median feature subset performance for various training and testing regimes. Features: C=color, T=texture, L=laser. Data sets: S=14 individual images; A=all-image digest (first letter=training, second=testing).

comprising these best individual performers (color and texture, texture and laser, etc.) were then trained, with the results shown in the **SS** column of Table 1.

As a baseline for performance assessment, the median proportion of feature vectors labeled "road" over each of the the 14 pairs was 48.8%. For the portion of feature vectors containing laser information, this road fraction was higher: 59.9% for the 15×15 subimages and 57.0% for the 31×31 ones.

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 $\mathbf{D}$ was created from the set of 107 images by randomly selecting 5% of each image's feature vectors and concatenating them. Training was performed on $\mathbf{D}$ for the four combined feature sets exactly as if it were a larger version of an image $\mathbf{I}_i$, yielding good results which are shown in the **AA** column of the table. The poor generality of the single-image models learned in the previous subsection is demonstrated by testing them on $\mathbf{D}$; the median performance over the 14 images is given in column **SA** of the table. Accuracy drops dramatically because of the presentation of road and background types not seen in the single image training.

The representative fidelity of using the digest for training can be seen in the similarity of the scores obtained in column **AS** by training on the digest and testing on the individual images to those achieved by training and testing on the digest alone (**AA**). However, the performance declines slightly for each feature set when switching training from individual images to $\mathbf{D}$. This is likely because the variety of road and background types in $\mathbf{D}$ cause a greater mixing of road/non-road points in feature space, necessarily increasing the error of any decision surface.

of computing the transformation did not seem to be worth the mixed results.

One model per road type Of $\mathbf{D}$'s 3424 feature vectors, 1619 or 47.3% were labeled "road." k-means clustering [13] was used to group road-labeled feature vectors in $\mathbf{D}$ ($k = 2, 3, 4, 5$) for the best color feature set, the best texture feature set, and the best color and texture feature set[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.

The road type of each of the 14 representative pairs was computed from the nearest cluster center to the mean of a small set of feature vectors assumed to be inside that image's road region. This set of 4×4 feature locations was defined by a square centered horizontally in the image and at its bottom (roughly $165 \leq x \leq 185, 180 \leq y \leq 210$). For a particular feature set and k, a digest $\mathbf{D}_i$ was made for every cluster by taking a nearly equal number of randomly selected feature vectors (including laser information ignored during the clustering process) from the images in it such that the size of $\mathbf{D}_i$ was 640. A separate neural network was trained for each cluster's digest.

Using just the best color feature set for clustering, for example, the median training accuracy for the color and laser combination over all of the clusters is 90.2% for $k = 2$, 94.5% for $k = 3$, 92.7% for $k = 4$, and 93.3% for $k = 5$. Compared with the model-per-image and one-model results in the second row of Table 1, this multi-type approach looks quite good. The performance surpasses that of fitting one model to all of the data, while still exhibiting the generality that training one model per image clearly lacks.

5 Conclusion

We presented a preliminary version of 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.

We have obtained encouraging results using support vector machines with a radial basis function kernel [9] as a road classifier, but did not have time to test them on all of the data. For completeness, we would also like to try a few other values for the number of

[4]The algorithm was run 50 times for each k and feature set; the result exhibiting the lowest within-cluster scatter to between-cluster scatter ratio was used.

neural network hidden units. Furthermore, the data set needs to be augmented to capture the visual and structural effects of temporal variations such as time of day, weather, and season for more generality.

References

[1] C. Taylor, J. Malik, and J. Weber, "A real-time approach to stereopsis and lane-finding," in *Proc. IEEE Intelligent Vehicles Symposium*, 1996.

[2] E. Dickmanns, "Vehicles capable of dynamic vision," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1997, pp. 1577–1592.

[3] D. Pomerleau, "RALPH: Rapidly adapting lateral position handler," in *Proc. IEEE Intelligent Vehicles Symposium*, 1995, pp. 506–511.

[4] S. Smith, "Integrated real-time motion segmentation and 3D interpretation," in *Proc. Int. Conf. Pattern Recognition*, 1996, pp. 49–55.

[5] J. Crisman and C. Thorpe, "Unscarf, a color vision system for the detection of unstructured roads," in *Proc. Int. Conf. Robotics and Automation*, 1991, pp. 2496–2501.

[6] T. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[7] M. Swain and D. Ballard, "Color indexing," *Int. J. Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[8] B. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

[9] N. Cristiani and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[10] J. Bouguet, "Camera Calibration Toolbox for Matlab," Available at `www.vision.caltech.edu/bouguetj/calib_doc`. Accessed May 11, 2001.

[11] M. Elstrom, P. Smith, and M. Abidi, "Stereo-based registration of LADAR and color imagery," in *SPIE Conf. on Intelligent Robots and Computer Vision*, 1998, pp. 343–354.

[12] H. Demuth and M. Beale, "Matlab Neural Network Toolbox User's Guide, Version 4.0," The MathWorks Inc., 2000.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification, 2nd ed.*, John Wiley and Sons, 2001.