Autonomous Learning for Long Range Vision in Mobile Robots

Raia Hadsell[1], Pierre Sermanet[1,2], Ayse Erkan[1], Koray Kavukcuoglu[1], Urs Muller[2], Yann LeCun[1]
(1) Courant Institute of Mathematical Sciences, New York University
(2) Net-Scale Technologies, Morganville, NJ 07751, USA

Designing a vision-based autonomous robot that can navigate through complicated outdoor environments is an extremely challenging problem that is far from solved. However, through use of a self-supervised realtime learning strategy and a trained deep belief net for compact feature representations, we have developed a long-range vision system that succeeds in accurately classifying obstacles and traversable areas that are 200 meters or more distant, bringing us closer to the goal of human-level autonomous driving.

Motivation: Shortcomings of Stereo Vision

The existing paradigm for vision-based mobile robots relies on hand-tuned heuristics: a stereo algorithm produces a (x, y, z) point cloud and traversability costs are assigned to points based on their proximity to a ground plane [2, 1]. However, stereo algorithms that run in realtime often produce costmaps that are short-range, sparse, and noisy. Our learning strategy uses these stereo labels for supervision to train a realtime classifier. The classifier then predicts the traversability of all visible areas, from close-range to the horizon. For accurate recognition of ground and obstacle categories, it is best to train on large, discriminative windows from the image, since larger windows give contextual information that is lacking in color and texture features. Other research has explored the use of online learning for mobile robots, but their methods have been largely restricted to simple color/texture correspondences [3, 4, 6].

Feature Extraction, Normalization, and Realtime Classification

The visual windows are relatively high dimensional (12x25x3 pixels), necessitating reduction to a concise feature representation. This is crucial in order to reduce processing time as well as remove statistical redundancies and extract meaningful features. With a good feature representation, an online learning algorithm can efficiently learn to discriminate different visual categories, using contextual information in the large windows. We use a deep belief network approach to feature extraction, training an autoencoder (offline) with unlabeled data from diverse outdoor log files [5] (see Figure 2). The network has 20 7x6 first layer filters and 300 6x5 second layer filters, producing a 100 dimension feature vector for each input. Max-pooling between the layers is done with a 1x4 kernel (no height pooling; 4x width pooling).

On every frame, the image is normalized, features are extracted, labels are assigned, the classifier is trained, and the entire image is classified. Normalization is done in 2 ways: by converting to YUV color space and doing contrast normalization on the Y channel, and by building a *scale-invariant image pyramid* (see Figure 1b). By subsampling the image at different locations, we can normalize the effects of distance on scale and thus have better generalization from near range to far range. A stereo module assigns labels to the processed windows that are close range (5 to 12 meters). The labels are in 5 categories: super-ground, ground, footline, obstacle, and super-obstacle (see Figure 1a). The classifier is then trained on these labeled samples, plus others from previous frames (the samples are collected in a ring buffer that both balances samples in the different categories and acts as a short-term memory). The classifier is a 5 class logistic regression, trained with stochastic gradient descent.

Evaluation

The long-range vision system has been implemented and tested using the LAGR (Learning Applied to Ground Robots) platform. The long range classifier allows the planner to avoid dead ends and navigate towards distant paths (see Figure 3). The long-range vision runs at 2-3 Hz, so it is processed in a separate thread from the main control loop, which runs at 10-15 Hz. This multiple-thread architecture allows the vehicle to nimbly avoid close obstacles while using the long-range vision for strategic long range planning. The full system has been thoroughly tested in the field, on courses containing tight paths, distant cul-de-sacs, complex obstacles, and varying ground types. Videos of the robot's performance on such courses will be shown at the workshop.

References

[1] S. B. Goldberg, M. Maimone, and L. Matthies. Stereo vision and robot navigation software for planetary exploration. March 2002. 1
[2] A. Kelly and A. Stentz. Stereo vision enhancements for low-cost outdoor autonomous vehicles. *ICRA Workshop WS-7*, May 1998. 1
[3] D. Kim, J. Sun, S. M. Oh, J. M. Rehg, and B. A. F. Traversibility classification using unsupervised on-lne visual learning for outdoor robot navigation. May 2006. 1
[4] R. Manduchi, A. Castano, A. Talukder, and L. Matthies. Obstacle detection and terrain classification for autonomous off-road navigation. *Autonomous Robot*, 18:81–102, 2003. 1
[5] M. Ranzato, Y. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *NIPS*, 2007. 1
[6] B. Sofman, E. Lin, J. Bagnell, N. Vandapel, and A. Stentz. Improving robot navigation through self-supervised online learning. In *Proc. of Robotics: Science and Systems (RSS)*, June 2006. 1

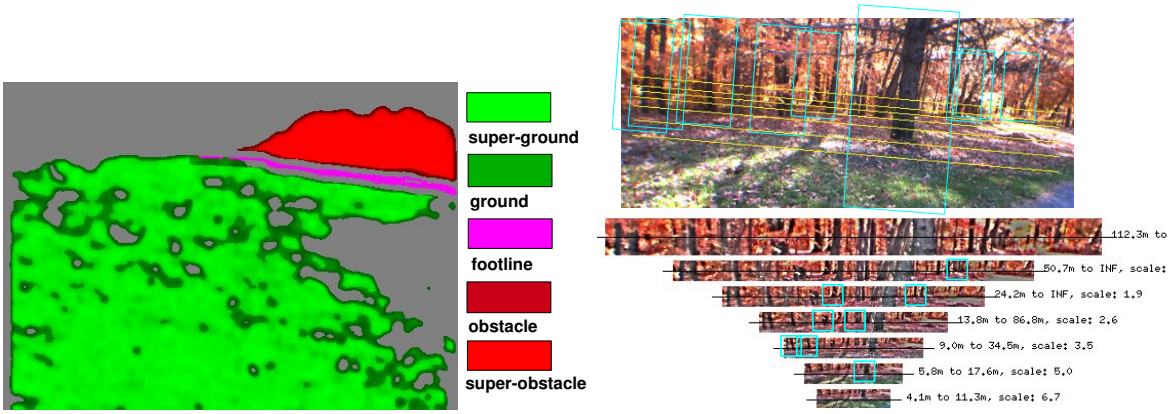

Figure 1: **Left**: The online classifier is trained with 5 categories: *super-ground* (ground fills the window, confidence is high), *ground* (ground visible in window, medium confidence), *footline* (footline of obstacle is centered in window), *obstacle* (obstacle is visible in window, medium confidence), *obstacle* (obstacle fills the window, high confidence). These training categories correspond well to natural visual categories, producing very good classification results. **Right**: The image is distance-normalized to produce a 7-level pyramid that provides scale-invariant windows for training.

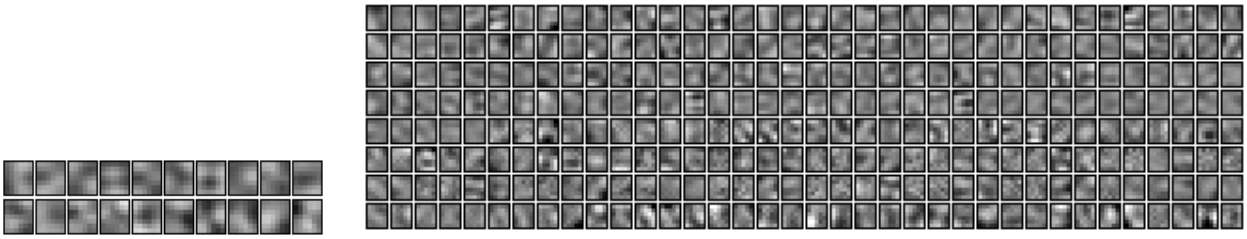

Figure 2: The filters learned by the autoencoder deep belief network. *Left*: First level kernels. *Right*: Second level kernels. The network is trained offline, unsupervised, using data from 150 logfiles in diverse environments. The input to the network is a 12x25x3 YUV window from the image, and the output is a 100 dimension feature vector.

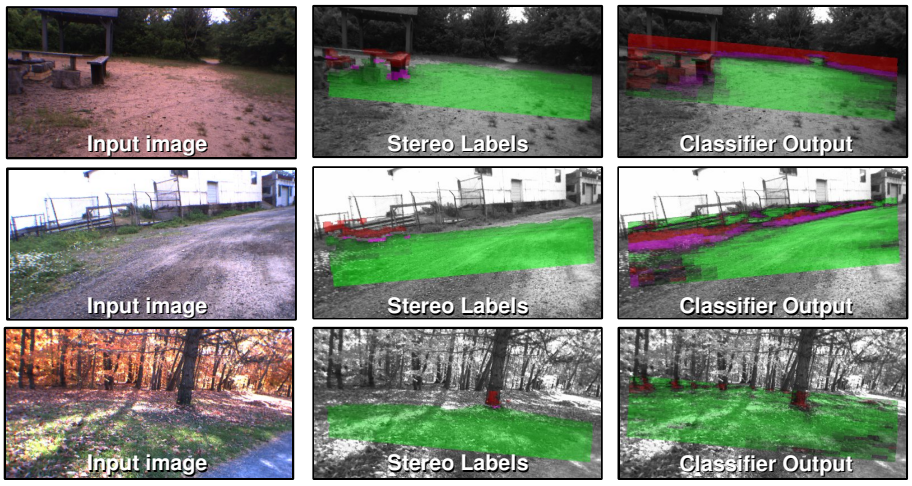

Figure 3: *Left* images are the RGB input image, *center* images show the training labels from the stereo module overlaid on BW input image (5 to 12 meters), and *right* images show the classifier labels (5 to approximately 200 meters). Green is Traversable, Pink is Footline, and Red is obstacle.