[8] A. Dattasharma and S. S. Keerthi, "An augmented Voronoi roadmap for 3D translational motion planning for a convex polyhedron moving amidst convex polyhedral obstacles," *Theoretical Computer Sci.*, vol. 140, no. 2, pp. 205–230, Apr. 1995.
[9] M. Frosky, M. Garber, M. Lin, and D. Manocha, "A Voronoi-based hybrid motion planner," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2001, pp. 55–60.
[10] K. Hoff, T. Culver, J. Key, M. Lin, and D. Manocha, "Interactive motion planning using hardware-accelerated computation of generalized Voronoi diagrams," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, pp. 1931–1937.
[11] L. Kavraki, P. Svestka, J.-C. Latombe, and M. Overmars, "Probabilistic roodmaps for path planning in high-dimensional configuration spaces," *IEEE Trans. Robot. Autom.*, vol. 12, no. 4, pp. 566–580, Aug. 1996.
[12] J. J. Kuffner and S. M. LaValle, "RRT-connect: An efficient approach to single-query path planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, pp. 995–1001.
[13] V. Lumelsky and A. Stepanov, "Path planning strategies for point mobile automaton moving amidst unknown obstacles of arbitrary shape," *Algorithmica*, vol. 2, pp. 403–430, 1987.
[14] M. Mason, K. Y. Goldberg, and R. H. Taylor, "Planning sequences of squeeze-grasps to orient and grasp polygonal objects," in *Proc. 7th CISM-IFToMM Symp. Theory, Practice Robots, Manipulators*, 1988, pp. 531–538.
[15] C. Ó'Dúnlaing, M. Sharir, and C. K. Yap, "Generalized Voronoi diagrams for moving a ladder. I: Topological analysis," *Commun. Pure Appl. Math.*, vol. 39, pp. 423–483, 1986.
[16] C. Ó'Dúnlaing and C. K. Yap, "A "retraction" method for planning the motion of a disc," *J. Algorithms*, vol. 6, pp. 104–111, 1985.
[17] C. Pisula, K. Hoff, M. Lin, and D. Manocha, "Randomized path planning for a rigid body based on hardware accelerated Voronoi sampling," in *Proc. Workshop Algorithmic Foundations Robot.*, 2000, pp. 279–292.
[18] N. S. V. Rao, S. Kareti, W. Shi, and S. S. Iyenagar, "Robot navigation in unknown terrains: Introductory survey of non-heuristic algorithms," Oak Ridge Nat. Lab., Oak Ridge, TN, Rep. ORNL/TM-12410:1-58, Jul. 1993.
[19] S. Rusaw, "Sensor-based motion planning in $SE(2)$ and $SE(3)$ via nonsmooth analysis," Univ. Oxford, Oxford, U.K., Tech. Rep. PRG-RR-01-13, Aug. 2001.
[20] S. A. Wilmarth, N. M. Amato, and P. F. Stiller, "Motion planning for a rigid body using random networks on the medial axis of the free space," in *Proc. 15th Annu. ACM Symp. Computat. Geom.*, Jun. 1999, pp. 173–180.
[21] Y. Yu and K. Gupta, "Sensor-based probabilistic roadmaps: experiments with an eye-in-hand system," *Adv. Robot.*, vol. 41, no. 6, pp. 515–536, 2000.

Robot Steering With Spectral Image Information

Christopher Ackerman and Laurent Itti

***Abstract*—We introduce a method for rapidly classifying visual scenes globally along a small number of navigationally relevant dimensions: depth of scene, presence of obstacles, path versus nonpath, and orientation of path. We show that the algorithm reliably classifies scenes in terms of these high-level features, based on global or coarsely localized spectral analysis analogous to early-stage biological vision. We use this analysis to implement a real-time visual navigational system on a mobile robot, trained online by a human operator. We demonstrate successful training and subsequent autonomous path following for two different outdoor environments, a running track and a concrete trail. Our success with this technique suggests a general applicability to autonomous robot navigation in a variety of environments.**

***Index Terms*—Autonomous robot, Fourier transform, gist of a scene, navigation, path following, vision.**

I. Introduction

Previous use of vision for robot navigation has often assumed a known or otherwise highly constrained environment [1]. In particular, successful autonomous indoor and outdoor navigation has been demonstrated with model-based approaches, where a robot compares a manually specified or learned geometric model of the world to its sensory inputs [2]. However, these approaches are limited by design to environments amenable to geometric modeling, and where reliable landmark points are available for model matching [3], [4].

Model-free or mapless algorithms have been proposed to address these limitations. In "view-based" mapless approaches, scene snapshots and associated motor commands are memorized along a route during training. During autonomous navigation, incoming images are matched against learned ones, to look up motor commands [5]. Images of the training environment may be stored explicitly and matched against new inputs by cross-correlation [6]. Effective in small environments, such approaches tend to generalize poorly (new scenes must correlate with learned ones), and require prohibitive memory and computation in larger environments. To alleviate these requirements, learned information may be implicitly stored in the weights of a trained neural network; typically, then, images are first reduced to a few landmark or characteristic regions, to reduce network complexity [7], [8]. Finally, computationally expensive algorithms involving much image processing and segmentation [9], including texture recognition [10], [11], stereo vision [12], or large rule sets [13] have been proposed to understand the environment. Although less explicit than model-based approaches, these algorithms also tend to incorporate a high degree of world knowledge and assumed environmental structure through design and tuning of feature detectors and rule bases. Here we take inspiration from biological vision in developing a mapless system with low memory, computation, and world-knowledge requirements.

The early stages of visual processing in the primate (including human) brain are believed, in a first approximation, to be organized in

Manuscript received March 22, 2004. This paper was recommended for publication by Associate Editor J. Kosecka and Editor S. Hutchinson upon evaluation of the reviewers' comments. This work was supported in part by the National Science Foundation, in part by the National Eye Institute, in part by the National Geospatial Intelligence Agency, and in part by the Zumberge Research Innovation Faculty Fund.

The authors are with the Computer Science Department, University of Southern California, Los Angeles, CA 90089 USA (e-mail: christopher.ackerman@gmail.com; itti@pollux.usc.edu).

Digital Object Identifier 10.1109/TRO.2004.837241

a triadic architecture recently described by Rensink [14] (also see [15] and [16]). The first stage consists of a massively parallel computation of simple visual features over the entire visual field. These features include intensity contrast, computed by neurons with center-surround receptive fields [17]; color opponency [18]; local orientation, detected by Gabor-like receptive fields [19]; stereo disparity [20]; motion energy [21]; and slightly more complex features including corners, T-junctions, and other proto-objects [22]. This stage feeds into two parallel processing branches. One is concerned with directing attention and gaze toward interesting locations, based on combined bottom-up (image-based) salience and top-down (voluntary) guidance, and with recognizing attended objects [15]. Persistent representations of a few attended and recognized objects are stored in working memory for later use [23], such as manipulation or landmark-based navigation. The second branch is concerned with the very rapid, most probably feedforward, coarse identification of the global setting of the scene, including its nature or "gist" (e.g., indoor versus outdoor, city versus beach) and approximate spatial layout [23]–[25]. We have previously developed an algorithm for bottom-up, saliency-based guidance of focal visual attention [26], [27]. It has been widely applied to the automatic detection of attention-grabbing objects in unconstrained static and dynamic scenes [15], [28], [29], and has been shown to highly significantly correlate with human eye movements [30], [31]. During navigation, this attentional branch is useful for rapid detection and identification of landmarks and unexpected spatially localized events [32], [33]. Here we hypothesize that the gist/layout branch provides information useful for basic navigation, such as presence of borders and degree of obstruction.

Recent work by Torralba and Oliva [34]–[36] classifies scenes according to their spectral components. Using only global or coarsely localized spectral information to capture the overall structure of a scene, they can compute high-level scene descriptors, such as openness and naturalness, and by casting images into this low-dimensional space, they classify scenes into semantic categories like mountains or city. Our approach is similar, but rather than semantic categorization, we seek to extract information useful for a particular action, navigation. The Fourier amplitude spectrum reflects an image's dominant contours, their orientations, widths, and lengths, and this information should be able to describe the relevant features of a path. It also reflects the degree of high- versus low-frequency information in an image, which combined with the above can describe the potential for movement within the scene.

It is of interest to employ neurally plausible visual computations, because the mammalian brain exhibits unmatched performance at using visual information to guide behavior like navigation. There is experimental psychophysical evidence that humans can recognize the overall gist of scenes after very brief presentations, and this could be facilitated by the rapid processing of spatial frequency information. For example, humans can identify semantic scene type after 45–135 ms exposure and have the capacity for low-resolution (not enough to define objects) scene recognition [37]. Rapid high-level scene classification does not require foveal vision [38] or color [39]. Rapid categorization of images for the presence or absence of certain classes of objects does not require attention [40], supporting the idea of a separate processing branch operating in parallel with attentional vision. As noted by [41], it is reasonable to think that global gist recognition may build on spatial-frequency-selective neurons in the primary visual cortex, as evolutionarily early mammals lack any higher visual area.

Using only global or very coarsely localized spatial frequency information to characterize the *"navigational gist"* of a scene, we apply these ideas to the task of path following. We demonstrate that scenes that vary along four navigational class descriptors have distinctive Fourier spectra, allowing a mapping from spectral information to scene class. Then we develop an architecture by which spectral information can be mapped to motor commands, via batch or online learning.

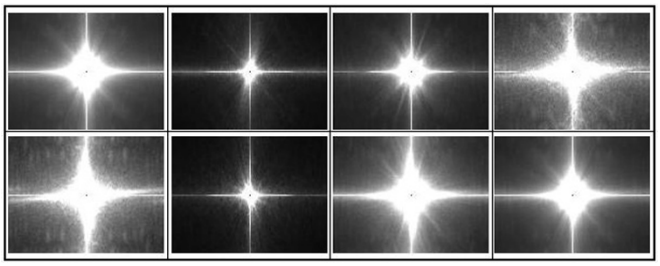

Fig. 1. Averaged Fourier amplitude spectra for all human-labeled images. From left to right: Path (top; $n = 912$) versus Nonpath (bottom; $n = 140$). Left (top; $n = 54$) versus Right (bottom; $n = 24$). Open (top; $n = 256$) versus Closed (bottom; $n = 302$). Crowded (top; $n = 96$) versus Empty (bottom; $n = 898$).

II. Classification for Navigation

A. Overview

We consider the following high-level information useful for path following. First, the direction in which the path borders are turning (for steering); second, whether the agent is on an open stretch or in a closed environment (which affects speed); third, the density of obstacles in the environment (to also modulate speed, and to trigger other behaviors like obstacle avoidance, for example, using an attentional vision system to locate and identify them); fourth and finally, recognizing whether the agent is on a path can determine whether the other features are to be used as above, or some other control behavior is necessary. Our system uses a trained feedforward neural network to estimate these four features from global or coarsely localized image spectral information. The results show that much intuitively navigationally useful information can be extracted from such computationally inexpensive analysis.

That there is sufficient information in the Fourier amplitude spectra of the luminance of the images to make the discriminations described above is supported by the averaged spectra of iconic images of each feature extreme, as illustrated in Fig. 1. All 1630 photographs used to create Fig. 1 were taken on the University of Southern California (USC) campus at slightly above ground level, at a robot's eye view (see next section). Thus, they are dominated by path borders, if present, and other potential ground-level obstacles or barriers. Paths are characterized by prominent contours, corresponding to path borders, extending off into the distance at various angles. As shown in Fig. 1, left, this is reflected in the averaged amplitude spectrum of path scenes by higher activity in the diagonal orientations, compared with the vertical and horizontal axes; in contrast, the nonpath scenes are more dominated by vertical and horizontal orientations, and also have more high-frequency content, reflecting the relative smoothness of the path portion of path scenes, compared with nonpath scenes. The direction along which the path is oriented is also reflected in the spectra. As shown in Fig. 1, middle left, leftward-oriented scenes have increased activity in the first quadrant (i.e., the path is at $\sim 135^\circ$; the borders are oriented so that they appear in the intensity gradient as a wave oriented at $\sim 45^\circ$). Rightward-oriented scenes show the complementary pattern. Open scenes, with extended depth and low horizon, have less high-frequency energy than closed scenes, whose close-up barriers presumably supply high-frequency details not visible on the more distant borders in open scenes (Fig. 1, middle right). Scenes crowded with obstacles contain more high-frequency energy than empty scenes, for reasons similar to the above, and are somewhat biased in the horizontal frequency direction (vertical lines in image space), reflecting vertical obstacles such as people (Fig. 1, right).

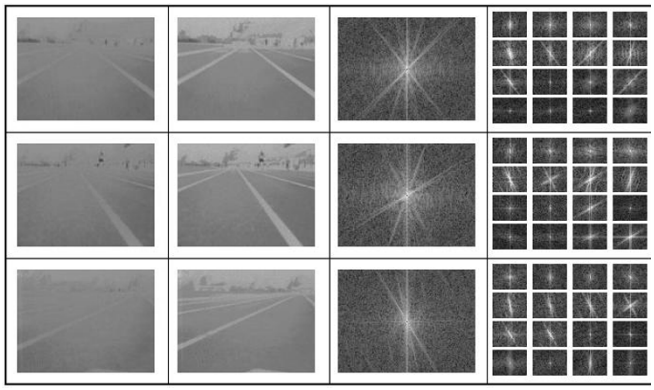

Fig. 2. Steps in transformation from RGB image to Fourier amplitude spectra. Left to right: original image, brightness-only and normalized image, amplitude spectrum from image Fourier transform, and amplitude spectra from transform of each of 16 nonoverlapping 30×40-pixel image tiles. Shown are three Beobot-captured running track images. Top to bottom: straight, leftward turning, and rightward-turning image class exemplars.

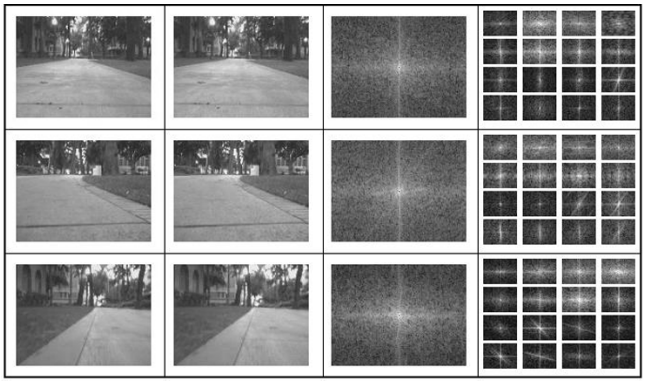

Fig. 3. Steps in transformation from RGB image to Fourier amplitude spectra. Format is identical to Fig. 2, but the environment is a campus trail.

B. Images

The robot from whose vantage point all images were captured and whose navigation is being guided was built on the chassis of a standard radio-controlled car, with an on-board Beowulf cluster of four Pentium-III CPUs running Linux, and is referred to as a Beobot [42]. Our algorithm only used one of the available CPUs, leaving ample computational resources for other tasks, including attention and object recognition.

To train and test our algorithm, we collected 1343 red–green–blue (RGB) images taken from a Beobot-mounted camera. Images were labeled by one human along the dimensions of left-right orientation, depth, obstacles, and as path, nonpath, or ambiguous. For orientation, each image was assigned to one of eight different classes, from far left to far right. For depth and obstacles, each image was assigned to one of six classes, from closed to open, and from empty to crowded, respectively. Of these, 896 images were used for training and 447 for testing. The images all were photographs of daytime outdoor scenes taken on the USC campus, especially but not exclusively along various paths, and on a running track. It is important to note that nothing in our technique is specifically designed or tuned to the class of images used here for testing. Presumably, similar results could be obtained in very different environments, such as indoors. Pictures were 120×160 or 240×320 pixels; the latter were downscaled to 120×160. Images varied not only by location and resolution, but also time of day and cloud cover, and thus overall illumination and shadows, and camera angle. Figs. 2 and 3 show sample images.

C. Procedure

RGB images were converted to hue, saturation, and value (HSV), with only the value component being retained. Each image was normalized for luminance by subtracting its mean and dividing by its standard deviation. Next, each image was discrete Fourier transformed. Figs. 2 and 3 show the image-transformation process for images of varying orientations and environments. The Fourier amplitude spectrum is a $120 \times (160/2 + 1) = 9720$-element array of real numbers. To reduce dimensionality to make learning tractable, we consider aggregate responses over 40 localized masks applied to the Fourier spectra. The masks were chosen such that they are equivalent to convolving the original image with 40 log-Gabor filters of varying scales and orientations, in a manner similar to that recently described in [35] and [36]. The filters are tuned to five spatial frequencies, with eight different orientations at each frequency. The pointwise products between the Fourier amplitude spectrum and weighing windows corresponding to the filters are computed, and the resulting 40 scalars are saved as a feature vector (FV), such that

$$\mathrm{FV}_i = \sum_{x,y=0}^{N-1} A(fx, fy) Gi(fx, fy) \tag{1}$$

where $A(f_x, f_y)$ is the amplitude spectrum, and G_i are the filters.

There are intuitive and experimental [43], [44] reasons to believe that spatial relationships between structural scene components may provide finer discrimination along our dimensions. To see if capturing these by including coarsely localized information would benefit classification, we next performed the above procedures on 4×4 nonoverlapping 30×40-pixel image tiles. Such tiles are still large enough that they hold macro-level features of the scene; as in the global case, we are capturing gross structural elements and not fine object contours or textures. As this would yield $16 \times 40 = 640$ features, to reduce this to a more manageable number, we retained only the first 40 principal components (PC) of the features computed over the training set. We apply the PC to the log-Gabor filters [45], producing new filter sets, and then apply these to the Fourier spectra. These are summed over all tiles for each image to produce a scene-level descriptor. Each of the 40 global feature inputs are thus computed by

$$\mathrm{PCFV}_i = \sum_{j=1}^{16} \sum_{x,y=0}^{N-1} A_j(fx, fy) \mathrm{PCG}_{i,j}(fx, fy) \tag{2}$$

where $A_j(f_x, f_y)$ are the image tile spectra, and $\mathrm{PCG}_{i,j}$ are the PC log-Gabor filters. Fig. 4 shows the effect of this transformation, before outer (tile) summation, for 16 PC, in the image domain.

FVs for the 896 training images were used to train a fully connected, feedforward neural network using backpropagation. Eight output units encoded the eight orientation classifications, with the unit with highest activation being the winner, and one output unit each encoded the other three dimensions by their strength of activation.

D. Classification Results

Table I shows classification results of global and tiled models for each of the four dimensions. Two points should be noted. First, the low percentages in the "mistakes" row indicate that serious misjudgments, for example, mistaking a rightward orientation for leftward, were uncommon for both models. Second, spatial distribution information captured by using image tiles substantially helped orientation discrimination and obstacle detection. Several different network architectures, as well as several simpler linear classifiers, all yielded similarly superior results for the coarsely localized model over the global one. We therefore used this model for navigation.

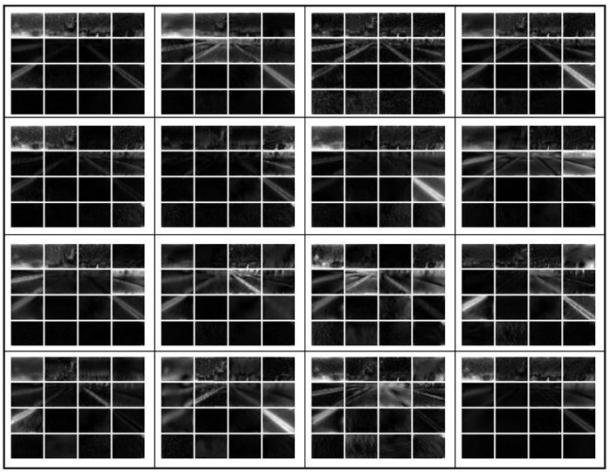

Fig. 4. Altered filters are applied to tiled track image spectra from Fig. 2, top, and reverted to image domain by the inverse Fourier transform to show, out of the many local features, textures, and clutter present in each tile, the image features (orientations, scales) that each of the first 16 PC is responsive to.

TABLE I
CLASSIFICATION RESULTS

	Pathness	Orientation	Depth	Obstacles
Accuracy, %	79.6/83.7	60.6/70.7	59.3/55.7	47.7/57.3
MSE	0.22/0.18	1.23/0.78	0.60/0.59	1.14/0.68
Mistakes, %	2.85/2.28	1.46/1.94	1.62/1.08	2.39/0.60

Global/Tiled. MSE = mean squared error. Mistakes: Pathness, excludes human-labeled ambiguous scenes; Orientation, incudes human-labeled left or right turning of any magnitude; Depth, Obstacles, includes human-labeled top (open, crowded) and bottom (closed, empty) thirds.

III. NAVIGATION

Having demonstrated that navigationally useful information can be extracted from image spectral components, we developed a preliminary system to demonstrate the use of this information in an autonomous robot navigation task. Orientation judgments relate straightforwardly to steering; we thus chose to use this information to test our approach as applied to path following. To enable continuous improvement through online training, we use a neural network that maps the processed visual features directly onto steering commands, and learns from its errors in real time. The network is a three-layer backpropagation perceptron with 40 inputs and 13 outputs, ranging from hard left to hard right, probabilistically chosen at each iteration based on strength of response. Initial network weights may be learned offline from images and steering commands captured with the robot under human control, or they may be random. Subsequently, weights are refined in an online manner, as a human operator may provide corrective action (via a remote control) to the steering choices of the autonomous algorithm. These human-issued commands are treated as target values, from which an error is computed for backpropagation.

Two common limitations of training a neural network by human operation are that humans are very efficient at maintaining accurate control, so that the learner may never be exposed to extreme situations, and that often the distribution of control commands is highly skewed (e.g., many straight path segments for a few turns), yielding unbalanced training and overrepresentation in the network of common situations compared with rare ones [46]. We minimize these potential hazards by letting the robot drive autonomously from the very beginning. Thus, instead of using every frame and the corresponding human control values for training, we only train on those frames during which the human operator is providing a nonzero corrective input. Instead of driving the robot as well as possible, the task of our operator is to wait until the robot starts heading off-track, and only then to apply brief corrective control. As it takes some time for the operator to realize that the robot is deviating from a desirable trajectory, and as the operator may voluntarily decide to wait until the last moment to apply correction, this approach ensures that extreme situations will be encountered during training. Also, because training only occurs in a need-based manner, rather than at every frame, it is not the potentially skewed distribution of all inputs, but the hopefully less-skewed distribution of corrected inputs, that influences the weights of the network.

We have allowed the robot to run autonomously using this method on both a running track and a paved concrete campus trail, with success. Qualitatively, the robot performed very well, even after limited training. For example, starting with random weights (entirely novel environment), the robot was able to run for extended time periods around the running track, after human corrections had been applied only during the first half of the track's circumference. Some generalization ability was also observed, as the robot required substantially fewer human corrections on its way back from finish to start of the trail, after training once on the way from start to finish. Future experiments will test for broader generalization, e.g., training on one path in the summer and running on another during winter, or training indoors then running outdoors. Videos of the Beobot in action are available online [47].

IV. DISCUSSION

An important contribution of our work is the finding that the classification of pathness, orientation, depth, and obstacles by Fourier amplitude spectra was robust, with very few labelings that were unambiguously wrong. This reinforces the idea that the type of low-level processing used here, originally motivated by previous work with semantic scene classification [35], [36], may be a generally applicable approach for the low-dimensional, global description of scenes along a variety of dimensions. This is particularly interesting in that the low-level visual features used in our system, based on oriented spatial frequency-bandpass filters, are compatible with the response properties of early visual neurons in the monkey brain, as characterized by physiological and other techniques [48].

Including rough spatial layout information is helpful in more complex environments, such as along campus trails. As can be seen in Fig. 2, orientation information is available in both the global amplitude spectrum and the coarsely localized spectra for the track environment, where the broad lines are prominent features. In the trail environment, however, as in Fig. 3, it is difficult to discern much from the global spectrum, which takes into account trees, buildings, and other objects. In the coarsely localized spectra, however, some of the oriented features that indicate path borders in the track environment of Fig. 2 are also found in the trail environment of Fig. 3.

Our navigation system, learning steering commands directly from spectral information, is attractive because it allows one to rapidly train the algorithm in novel environments, with a minimal amount of supervision, and adjust for mistakes over time. This holds promise for use in other navigation tasks, including a fuller implementation of path following, e.g., adding obstacle avoidance and landmark-based orienting behaviors exploiting focal attention and localized object recognition, or speed modulation based on scene depth. The current decision system is very simple, using a single neural network with no temporal memory (reflex agent), as our main focus is on feature extraction; it is thus remarkable that it achieved good real-world outdoor navigation performance. We avoid computationally expensive and time-consuming tasks such as segmentation or object recognition, so as to enable real-time operation at relatively high speeds (~ 5 mph, limited by safety), nor are we dependent on specific environmental features. Yet, we have demonstrated a working system on a physical robot. Although generalization

of our approach to arbitrary control tasks remains to be proven, in the task domain studied here, the system appears robust, flexible, and computationally efficient enough to run in real time on one Pentium-III processor.

ACKNOWLEDGMENT

The authors thank A. Torralba and A. Oliva for image processing Matlab code.

REFERENCES

[1] G. N. DeSouza and A. C. Kak, "Vision for mobile robot navigation: A survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 237–267, Feb. 2002.

[2] A. Kosaka and A. C. Kak, "Fast vision-guided mobile robot navigation using model-based reasoning and prediction of uncertainties," *Computer Vision, Graphics, Image Process.—Image Understanding*, vol. 56, no. 3, pp. 271–329, 1992.

[3] E. D. Dickmanns and B. Mysliwetz, "Recursive 3-D road and relative egostatic recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 199–213, Feb. 1992.

[4] V. Graefe and K. Kuhnert, "Vision-based autonomous road vehicles," in *Vision-Based Vehicle Guidance*, I. Masaki, Ed. New York: Springer, 1992, pp. 1–29.

[5] Y. Matsumoto, M. Inaba, and H. Inoue, "View-based approach to robot navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2000, pp. 1702–1708.

[6] S. D. Jones *et al.*, "Appearance-based processes for visual navigation," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 1997, pp. 551–557.

[7] P. Gaussier *et al.*, "Visual navigation in an open environment without map," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 1997, pp. 545–550.

[8] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-89-107, 1989.

[9] A. Gilg and G. Schmidt, "Landmark-oriented visual navigation of a mobile robot," *IEEE Trans. Ind. Electron.*, vol. 41, pp. 392–397, Aug. 1994.

[10] M. Rosenblum and L. S. Davis, "An improved radial basis function network for visual autonomous road following," *IEEE Trans. Neural Netw.*, vol. 7, pp. 1111–1120, Sep. 1996.

[11] C. Thorpe *et al.*, "Vision and navigation for the Carnegie-Mellon Navlab," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 362–372, May 1988.

[12] D. R. Murray and J. J. Little, "Interpreting stereo vision for a mobile robot," *Auton. Robots*, vol. 8, no. 2, pp. 161–171, 2000.

[13] M. A. Turk *et al.*, "VITS—A vision system for autonomous land vehicle navigation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 342–361, May 1988.

[14] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cogn.*, vol. 7, pp. 17–42, 2000.

[15] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[16] L. Itti, "Modeling primate visual attention," in *Computational Neuroscience: A Comprehensive Approach*, J. Feng, Ed. Boca Raton, FL: CRC Press, 2003, pp. 635–655.

[17] S. W. Kuffler, "Discharge patterns and functional organization of the mammalian retina," *J. Physiol.*, vol. 16, pp. 37–68, 1953.

[18] S. Engel *et al.*, "Color tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, pp. 68–71, 1997.

[19] J. G. Daugman, "Spatial visual channels in the Fourier plane," *Vis. Res.*, vol. 24, pp. 891–910, 1984.

[20] G. F. Poggio and B. Fischer, "Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey," *J. Neurophysiol.*, vol. 40, pp. 1392–1405, 1977.

[21] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A*, vol. 2, no. 2, pp. 284–299, 1985.

[22] A. Pasupathy and C. E. Connor, "Responses to contour features in macaque area v4," *J. Neurophysiol.*, vol. 82, pp. 2490–2502, 1999.

[23] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Annu. Rev. Psychol.*, vol. 50, pp. 243–271, 1999.

[24] A. Oliva and P. G. Schyns, "Diagnostic color blobs mediate scene recognition," *Cogn. Psychol.*, vol. 41, pp. 176–210, 2000.

[25] B. W. Tatler, I. D. Gilchrist, and J. Rusted, "The time course of abstract visual representation," *Perception*, vol. 32, pp. 579–592, 2003.

[26] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1254–1259, Nov. 1998.

[27] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, no. 10–12, pp. 1489–1506, 2000.

[28] ——, "Feature combination strategies for saliency-based visual attention systems," *J. Elect. Imag.*, vol. 10, no. 1, pp. 161–169, 2001.

[29] F. Miau and L. Itti, "A neural model combining attentional orienting to object recognition: Preliminary explorations on the interplay between where and what," in *Proc. IEEE Eng. Med. Biol. Soc.*, Istanbul, Turkey, Oct. 2001, pp. 789–792.

[30] D. Parkhurst *et al.*, "Modeling the role of salience in the allocation of visual selective attention," *Vis. Res.*, vol. 42, no. 1, pp. 107–123, 2002.

[31] R. J. Peters *et al.*, "Eye movements are influenced by short-range interactions among orientation channels," in *Proc. Soc. Neurosci. Annu. Meeting*, Nov. 2002, p. 715.12.

[32] M. F. Land and D. N. Lee, "Where we look when we steer," *Nature*, vol. 369, no. 6483, pp. 742–744, Jun. 1994.

[33] H. Shinoda, M. M. Hayhoe, and A. Shrivastava, "What controls attention in natural environments?," *Vis. Res.*, vol. 41, no. 25–26, pp. 3535–3345, 2001.

[34] A. Oliva, A. B. Torralba, A. Guerin-Dugue, and J. Herault, "Global semantic classification of scenes using power spectrum templates," in *Proceedings of Challenge of Image Retrieval Electronic Workshops in Computing Series*. Newcastle, U.K.: Springer-Verlag, 1999.

[35] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[36] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Amer. A Opt. Image Sci. Vis.*, vol. 20, no. 7, pp. 1407–1418, Jul. 2003.

[37] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition," *Psychol. Sci.*, vol. 5, pp. 195–200, 1994.

[38] M. Fabre-Thorpe *et al.*, "Rapid categorization of extrafoveal natural images: Implications for biological models," in *Computational Neuroscience: Trends in Research*, J. Bower *et al.*, Ed. New York: Plenum, 1998, pp. 7–12.

[39] A. Delorme *et al.*, "Rapid categorization of natural scenes is color blind: A study in monkeys and humans," *Vis. Res.*, vol. 40, pp. 2187–2200, 2000.

[40] F. F. Li *et al.*, "Rapid natural scene categorization in the near absence of attention," in *Proc. Nat. Acad. Sci.*, vol. 99, 2002, pp. 9596–9601.

[41] C. Rasche and C. Koch, "Recognizing the gist of a visual scene: Possible perceptual and neural mechanisms," *Neurocomputing*, vol. 44, pp. 979–984, 2002.

[42] L. Itti *et al.*. (2003) Towards visually-guided neuromorphic robots: Welcome to the Beobot project. Beobot Team, USC, Los Angeles, CA. [Online]. Available: http://iLab.usc.edu/beobots/

[43] P. Lipson, E. Grimson, and P. Sinha, "Configuration based scene classification and image indexing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 1007–1013.

[44] A. Torralba and A. Oliva, "Scene organization using discriminant structural templates," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, pp. 1253–1258.

[45] J. G. Daugman, "Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, pp. 1160–1169, 1985.

[46] D. Pomerleau, "Efficient training of artificial neural networks for autonomous navigation," *Neural Computat.*, vol. 3, pp. 88–97, 1991.

[47] L. Itti *et al.*. (2003) Towards visually-guided neuromorphic robots: Beobots in action. Beobot Team, USC, Los Angeles, CA. [Online]. Available: http://iLab.usc.edu/beobots/gallery-action.shtml

[48] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.