Gist: A Mobile Robotics Application of Context-Based Vision in Outdoor Environment

Christian Siagian

Department of Computer Science
University of Southern California
Los Angeles, CA 90089

Laurent Itti

Department of Computer Science
University of Southern California
Los Angeles, CA 90089

Abstract

We present context-based scene recognition for mobile robotics applications. Our classifier is able to differentiate outdoor scenes without temporal filtering relatively well from a variety of locations at a college campus using a set of features that together capture the "gist" of the scene. We compare the classification accuracy of a set of scenes from 1551 frames filmed outdoors along a path and dividing them to four and twelve different legs while obtaining a classification rate of 67.96 percent and 48.61 percent, respectively. We also tested the scalability of the features by comparing the classification results from the previous scenes with four legs with a longer path with eleven legs while obtaining a classification rate of 55.08 percent. In the end, some ideas are put forth to improve the theoretical strength of the gist features.

1. Introduction

The field of mobile robotics, which hinges on solving tasks such as localization, mapping, and navigation, can be summarized into one central question: Where are we? This fundamental problem can be approached from several angles. A significant number of mobile robotics implementations use sonar, laser, or other range sensors [3, 7, 23]. In the outdoors these sensors become less effective in solving data association problem because, unlike the indoor environment, the existence of spatial simplifications such as flat walls and narrow corridors cannot be assumed. Contrasting with the regularity of indoor scenes, the surface shape of the outdoors varies tremendously, especially when considering environments with different terrains. It is very hard to predict the sensor input given all the protrusions and surface irregularities[8]. A slight pose change can result in a large jump in range reading because of tree trunks, moving branches and leaves. In addition, a flat surface for a robot to travel is an absolute must, otherwise we introduce a third (and mostly unaccountable) dimension.

Another way to obtain navigational input is by using the primary sensory modality in human: vision. Within the Computer Vision field, a large portion of approach towards scene recognition is object-based [1, 24, 9]. That is, a physical location is recognized by identifying surrounding objects (and their configuration) as landmarks. This approach involves intermediate steps such as segmentation, feature grouping, and object detection. The last step can be complex because the necessity of object matching from multiple views [11, 10]. It should be pointed out that this approach is mostly used indoors for the simplicity of selecting a small set of anchor objects that keep re-occuring in the environment. In the generally spacious outdoors, objects tend to be farther away from each other. Because of inherently noisy camera data - objects are more obscured and smaller in comparison to image size - such layered approach produces errors that are carried over and amplified along the stream of processing.

A different set of approaches within the Vision domain looks for regions and their relationship in an image as a signature of a location. Katsura [6] utilize top-down knowledge for recognizing the regions such as the sky being at the top part of an image and buildings at the east and west of the frame. A scalability problem arises from the inability to chracterize the region more than simply its centroid. Matsumoto [10] uses template-matching to recognize a set of regions and, to a degree, encapsulates a richer set of features. However, the need for pixel-wise comparison may be too specific to allow flexibility of different views of the same location. Murrieta-Cid [11] goes a step further in that it uses the regions, particularly isolated-blob region, as an intermediate step to locate predetermined set of anchor landmarks. Here, there is a need for robust segmentation phase as well as identification of landmarks from multiple views.

The context-based approach, on the other hand, bypasses the traditional processing steps. Instead, it analyzes the scene as a whole and extract a low-dimensional signature for it. This signature in the form of a scene-level feature vector embodies the cognitive psychologists' idea of the

"gist" of a scene [17]. The hypothesis at the foundation of this technique is that it should produce a more robust solution, because random noise that locally may be catastrophic tends to average out globally. By identifying scenes, and not objects, we do not have to deal with noise in isolated regions. Our approach, which is biologically inspired, mimics the ability of human vision to collect coarse yet concise contextual information about an image in a short amount of time [16, 27, 22]. This gist information includes a rough spatial layout [12], but often lacks in fine-scale detail about specific objects [13]. Gist can be identified in as little as 45 - 135 ms [19], faster than a single saccadic eye movement. Such quick turnaround time is remarkable considering that it extracts quintessential characteristics of an image which can be useful for tasks such as semantic scene categorization (e.g., indoors vs. outdoors; beach vs. city), scale selection, region priming and layout recognition [25].

The challenge to discover a compact and holistic representation has been a research of various works. Renniger [18] and Malik use a set of texture information and keep track of them using a histogram to create an overall profile of an image. Ulrich and Nourbakhsh [28] build color histogram and perform matching using a voting procedure. Although an approach by Takeuchi [21] is meant to look for red buildings, the actual implementation uses a histogram of red texture pixels. In contrast to the previous approaches, which does not encode spatial information, Oliva [15] and Torralba performs Fourier Transform analysis to individual sub-region divided using a regularly-spaced grid, which then is correlated to several scene categories. Later on Torralba [26] uses steerable wavelet pyramid in the same manner. The core of our present research focuses on a process of extracting the gist features of an image from several domains that do not focus on specific locations of the image but still take into account a coarse spatial information.

2. Design and Implementation

Part of our contribution is that we present a more biologically plausible model which utilizes the rich image features from the visual cortex. The gist model is built using the Vision toolkit developed by Itti *at al.* [4] which feature the Saliency model and can be freely downloaded on the web. In the Saliency model, the image is processed through a number of low-level visual "channels" at multiple spatial scales. Within each channel, the model performs a center-surround operations between different scales to detect conspicuous regions for that channel. These results then is linearly combined to yield a saliency map. One of the goals here is to re-use the same intermediate maps, so that gist is computed almost for free once we have all the maps for attention. Our gist model make use of the already available orientation, color and intensity opponency features. We incorporate information from the orientation channel, which uses Gabor filters, at four different angles and at four spatial scales for a subtotal of sixteen sub-channels. The color channel (in two color opponency: red-green and blue-yellow) is composed of twelve different center-surround scale combination while the intensity channel (dark-bright opponency) is composed of six different center-surround scale combinations. That is a total of thirty-four sub-channels altogether. The following image 1 illustrates the features used by the gist model.

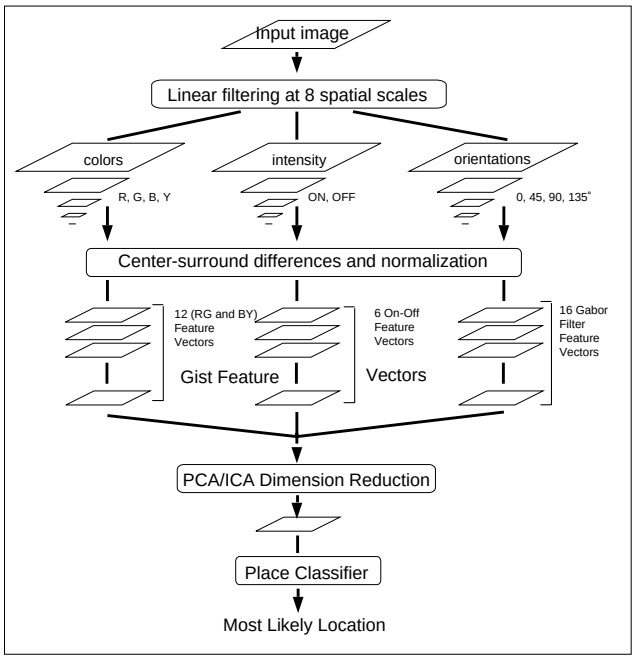

Figure 1: Visual Feature Channel Used in the Gist Model

For each of the thirty-four sub-channels, we extract the corresponding gist vector of 21 features from its filter output/feature map. We hypothesize that, because of the speed in which gist is calculated, they come from a series of simple feedforward operations. We thought that accumulation or averaging operations (as oppose to competition) are more plausible in finding general information of a set of values. Our 21 gist features can be visualized by a pyramid of subsection average-outputs (figure 2). The first, the tip of the pyramid, represents the mean of the entire feature map. The next four values are means of each of the quarter sub-region of the map - delineated by an evenly spaced two-by-two grid. The last sixteen values are the means of each of the sub-region delineated by a four-by-four grid. This approach is similar to Torralba [26] with the use of wavelet pyramid. The number of features could be increased by adding other domains such as motion and stereo, or using even smaller sub-regions (the next level would be an eight-by-eight grid: an additional 64 features) for more localized gist informa-

tion.

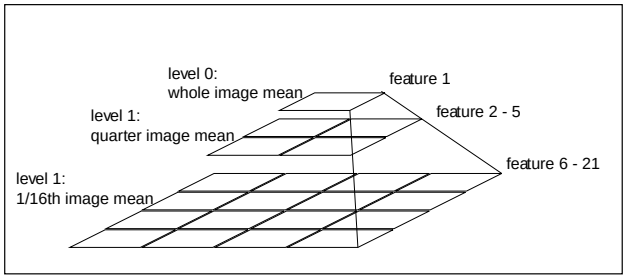

Figure 2: Pyramid of means of sub regions of an image

The current total number of feature dimension is 714, thirty-four (sub-channels) times twenty-one (features per sub-channels) features, which is a relatively high number for a lot of classification tasks. We reduce the feature dimension, using Principal Component Analysis (PCA) and then Independent Componenet Analysis (ICA) using FastICA [2], to a more practical number of 80 while still preserving the variance up to 99.8 percent for a set in the upwards of 2000 well-spaced images from the USC campus. Considering a virtually lossless reduction, it quite a surprise that our further PCA/ICA analysis does not yield observable pattern of redundancy that would have helped towards more principled feature reduction.

It is important to note that the algorithm has a brute force feel to it. Although more sophisticated gist computation should be incorporated, the current technique highlights the rapid nature of gist. The features themselves are hardly raw, as the low level processing produces contrast information in different domains and scales. Portions of information that seems to be omitted in the gist encoding are direct color distribution [28]. However, color distribution requires a sophisticated normalization process, especially in the outdoors where the light sources (sun and sky) can change in terms of luminance, chrominance, and direction. On the other hand, the center-surround features use of contrast strengthens the lighting invariance nature of the features, although they take out the absolute raw value of the image. Moreover, it can also construed as looking for edges surrounding regions but not the homogeneous regions (blobs) themselves. However, because the system uses a pyramid scale, it will pick up the region at a higher scale [13] and indirectly infer the absolute value information, still with the added lighting invariance.

3. Testing and Results

We test the technique at the USC campus. Figure 3 shows a highlighted path one would usually take to go to a bus stop in front of a building called “JEP house (JEP)” from the Computer Science department, the “Salvatori Building (SAL).” The path takes place in an outdoor environment that expands several city blocks. It requires navigation through the sidewalk, street crossing, and a much more open area in front of the school’s bookstore. We chose the path because they are quite typical of locations on campus. Refer to Figure 4 for the variety of the visual stimuli collected along the path. Although the scenes are different in terms of configuration the trees and buildings, they can be quite similar in appearance if we do not encode the region segmentation in detail. In addition, because of the overlapping trees and buildings happen quite often, the only reliable segmentation is simply sky and ground. It would take an expressive representation to provide a unique and robust segmentation signature.

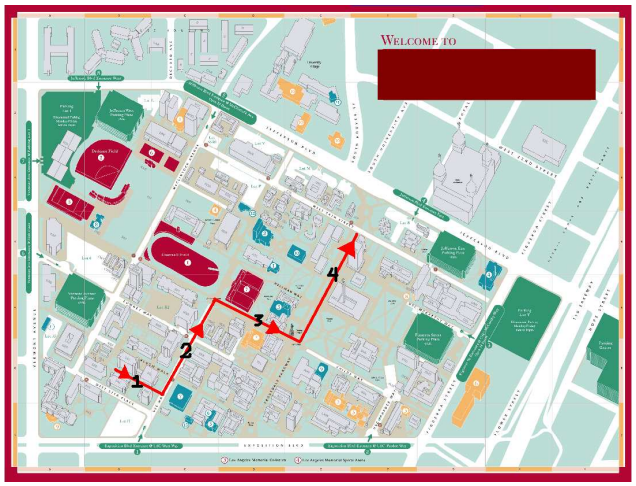

Figure 3: Map of USC with highlighted SAL to JEP route

To collect video data along the path, we use an 8mm handheld camcorder. The video clips are hardly stable. The camcorder is carried by a person who makes no attempt to smooth out the image jitter, the camera sometimes points up to the sky because of the bumpy road. At this point the data is still view specific, each location is only traversed from one direction. However, to perform a view invariant scene recognition, all we need to do is to train the system on multiple view images [26].

3.1. Experiment 1: Street Segment Classification

The first localization task is to assess the ability of the gist features to reliably encode path segments. Figure 3 shows that there are four segments on the SAL-to-JEP route. Each segment is a straight line and part of a long road with different looking path delineation as shown in figure 4. The path is divided this way because images within each segments share the majority of the background scenes. As a comparative study, we also classified a more spatially lo-

	Misclassification	% correct
Training		
Overall	23/1651	98.61%
Testing		
Overall	497/1551	67.96%
Segment 1	23/118	80.51%
Segment 2	311/575	45.82%
Segment 3	104/440	76.36%
Segment 4	59/418	85.89%

Table 1: Street Segment Classification

calized set of images in Experiment 2. Note that we do not remove scenes containing non-stationary objects (people, cars, bikes) or images dominated by trees from our data.

Figure 4: Sample images for each path segment of the SAL - JEP route

We perform the classification using a three-layer neural network, trained with back-propagation algorithm on the eighty PCA/ICA reduced gist features. The output layer has four nodes (the same as the number of segments) and we use an absolute encoding for training data. For example, if the correct answer is segment 1, the corresponding node is assigned 1.0, while the other is 0.0. This encoding allows for probabilistic ideal output for scenes in between places. For completeness, the intermediate layers have 250 and 100 nodes, respectively. Table 1 shows the result of this experiment.

After only taking small number of epochs to converge towards less than five percent error, table 1 shows that the network is able to classify the segments consistently during the testing phase, except for the second. The segment poses a bit of problem caused by the high number of people who walks on the busy road. The students as well as vehicles can move too close to the camera and occluded a large portion of the field of view. These fairly useless scenes may be removable by a bottom up approach using a motion field and segment out non-stationary objects that cannot be construed as reliable part of the scene. We also suggest a hardware solution to the problem: using a wide-angle lens (with correction calibration procedure to remove distortion artifact) so as to simulate peripheral vision perspective. Just as a comparison, if we take out the second segment, the correct classification jumps up to 80.94 percent. We also notice that however careful the filming is done, the camera can accidentally point toward the sky. We can consistently detect the event simply by noting there is very little activities on each channel.

3.2. Experiment 2: Building Segment Classification

We now compare Experiment 1 results with more difficult classification task: segmenting path based on locations where a particular building is viewable. This setup provides more accurate localization with each building span about 100 ft in length. Figure 5 labels each building segments. As with Experiment 1, we perform a multi-layer neural network based classification using back-propagation algorithm. At first we try to use the same network architecture (same number of input and intermediate nodes) as Experiment 1, but it is quickly become obvious that the network does not have the capacity to perform the classification. We increase the number of intermediate nodes from 250 and 100 to 400 and 200, respectively. The result shown at Table 2.

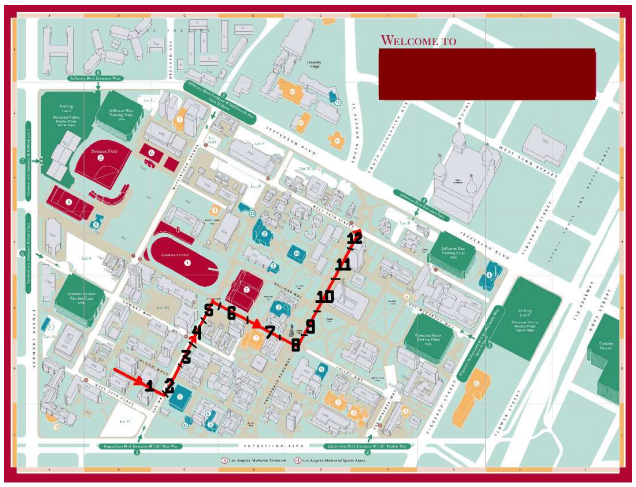

Figure 5: Building Segment of the SAL to JEP route

A quick glance at Table 2 reveals that the second classification is not as successful (below fifty percent). Taking a more in depth look at the testing phase, we discover some

	Misclassification	% correct
Training		
Overall	62/1651	96.24%
Testing		
Overall	797/1551	48.61%
Segment 1	14/112	87.50%
Segment 2	18/161	88.82%
Segment 3	104/168	38.10%
Segment 4	135/153	11.77%
Segment 5	75/105	28.57%
Segment 6	86/134	35.81%
Segment 7	69/200	65.50%
Segment 9	59/101	41.58%
Segment 10	11/43	74.42%
Segment 11	187/250	25.20%
Segment 12	39/125	68.80%

Table 2: Building Segment Classification

	Misclassification	% correct
Training		
Overall	96/5254	98.17%
Testing		
Overall	2139/4762	55.08%
Segment 1	28/118	76.27%
Segment 2	296/588	49.66%
Segment 3	321/436	26.38%
Segment 4	133/418	68.18%
Segment 5	178/211	35.97%
Segment 6	122/187	34.76%
Segment 7	61/95	35.79%
Segment 8	83/144	42.37%
Segment 9	426/1179	63.87%
Segment 10	201/1039	80.65%
Segment 11	142/331	57.10%

Table 3: Campus Street Classification

trends. At least one of the building-segments at a street segment has a high classification rate. For example: segments 2, 3, 4, and 5 are in the same street; building segment 2 is classified up to 87.50%. In several training and testing sessions, a higher percentage of the misclassifications hypothesize building-segments that are adjacent to the true segment. This would have suggest a possibility of significant overlapping scene features as a reason of misclassification. However, in other training sessions, such trend is not observed.

3.3. Experiment 3: Larger Street Segment Classification

Another comparison with experiment 1 is to scale up the number of scenes to classify. We increase the number of segments to eleven by constructing a longer route around the campus (oserve figure 6). As a point of reference, we keep most of the architecture of the neural network from experiment 2 except with one less output node. We recorded the result of our experiment in Table 3.

From table 3, we have mixed results. This is partially because of a large discrepancy between the training and testing run on several segments. A single run takes close to fifty minutes, enough time for the sun to change illumination. We run another filming the next day, however, the same problem occur again as the sun does not keep a constant illumination too long. We may need to perform a normalization in order to counteract the problem. An alternative would be to train the robot at several lighting condition, on multiple times of the day. A note, chance is at 9.1 percent when classifying data with eleven different classes.

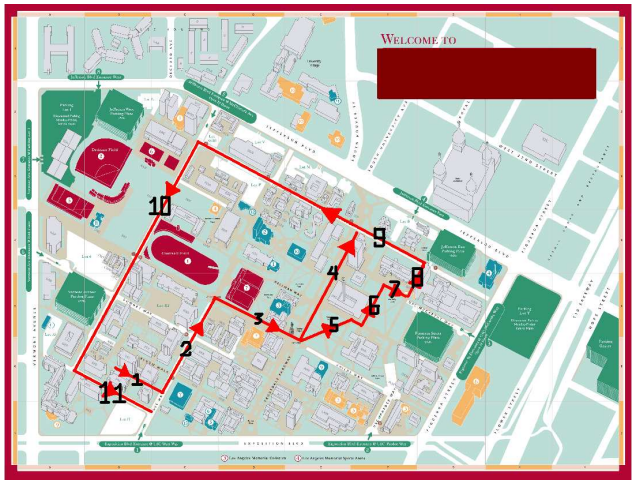

Figure 6: Longer Route Walking Around Campus

3.4. Experiment 4: Sub-channel Analysis

We also conduct an experiment to see whether there is an advantage in using information from a variety of feature domains, as we have done with orientation, color, and intensity channels. The argument put forth here is whether the notion of gist as spatially coarse cues applies to all domains or if there are only certain channels in which it is useful. This experiment will isolate each of the three channels, and individually train each for a given task. Previously it is mentioned that there are non-descriptive scenes that cannot be used for scene identification such as the ones dominated by trees or other featureless objects. We would like to know if the gist features are able to differentiate natural scenes and scenes containing man-made objects. The key distinction is that the natural scenes will then be construed as images

	Training		Testing	
	Misclass.	%Corr.	Misclass.	%Corr.
Combination	2/177	98.50%	16/120	86.67%
Orientation	5/177	97.17%	18/120	85.00%
Color	4/177	97.74%	33/120	72.50%
Intensity	4/177	97.74%	37/120	69.17%

Table 4: Natural vs. Man-Made Scenes Classification

full of trees while man-made scenes are scenes dominated by buildings. We use a fraction of database containing various semantic categories such as cities, farms, mountains, beaches, forests, and indoors for this task. Table 4 shows the results for the natural versus man-made scenes experiment.

From the results in Table 4 we discover that, for this task, the orientation channel (based on Gabor filters output in various scales and orientation) is the primary reason of the success. The orientation channel is able to generalize by making correct classification 85 percent of the time after the neural networks converge below five percent error during the training phase. On the other hand the intensity channel, which is based on the intensity center surround opponency, can only manage less than 70 percent classification after the network converge below 5 percent error during the training phase. The performance is marginally worse if we use the color channel because color is not as constraining a feature for the classification task.

It would be incorrect, however, to conclude that the color and intensity channels do not play a role in scene recognition task. It is possible that there are scene-related tasks in which these two channels will be valuable. Oliva [14] asserts that color will be of help if it has a diagnostic value to the scene. For example people are able to identify scenes of beaches faster with color as opposed to grey-scale images because beaches have a reliable layout structure with yellow sand and blue sea. It would be interesting to see if we implement an architecture in which these channels computes their gist features separately and competes for which hypothesis to use. Much like mixture of experts, the winner is the channel with the highest confidence level.

4. Discussion

We have shown that the gist features relatively succeed in classifying a large set of images without the help of temporal filtering which has been proven to reduce noise significantly [26]. In addition the system is asked to answer on each frame, despite its low confidence level. There are several factors that can be attributed to the performance. One of the strengths of the gist features is their non-reliance on specific locations, the features are computed from the whole image rather than smaller sub-regions, which increases stability over translational and rotational change of camera. Furthermore, a wide-angle lens as peripheral vision may help to see more of the scenes and less of the moving foreground objects because over a large set of images, dynamic foreground changes tend to averaged out. However, when we average out large spaces, some background details will be missed. This can be a cause for concern regarding the scalability of gist features considering degradation of results of experiment 2 and 3 with respect to experiment 1.

4.1. Compact layout representation

A way to increase the strength of the gist features, arguably, is to go to a finer grid to incorporate more spatial information. While it is an intermediate solution, the trade off may not be in our favor. If we go to next level in the pyramid (an eight-by-eight grid), the features is increased from 21 to 85 per sub-channel. It is not obvious what the gain is from the increase in spatial resolution. On one hand we get more resolution, on the other hand the features are more susceptible to noise that we want to avoid in the first place. We need to find a more expressive spatial decomposition (layout representation) that goes beyond the current grid setup.

4.2. Integration with Saliency Model

Another way to look at scalability problem is to accept the limitations of the gist features by themselves. When we visit a new city, if the buildings are roughly identical, every city block can look quite the same. We differentiate these locations by finding distinctive cues situated on the streets (post box, street signs) or buildings (name signs) which probably is lost when we average out the large sub-regions. There could be a way to incorporate these salient cues without exhaustive search (a drawback of object-based recognition), even with the help of the gist information.

Because our gist model is built under the vision toolkit [4], we can the incorporate results from its saliency model at minimum computational cost as both model uses the same Visual Cortex raw features. The saliency model has been shown to reliably predict which regions in an image attract visual attention of humans and other primates, as demonstrated by a high correlation between model saliency and actual human eye movements in [5]. From the point of view of desired results, gist and saliency appear to be opposites: most conspicuous locations (isolating them from other locations) versus general characteristics of an image by considering every region in the image in equal manner. Computationally, they are also opposites. In saliency the features from each location compete against each other (in level of conspicuousness). On the other hand, the gist module employs more of a cooperative operation in which the features are combined to produce an end result. These two models, when run in parallel, provide a more complete story

of the scene in question and have been successfully applied in various detection tasks[20, 25].

5. Conclusion

We have shown that context based vision can aid localization task. Because the raw gist features can be shared with other modules, it gives us the opportunity to attack the problem from multiple sides efficiently. In our case, robot localization from both the context and object-based perspective. The gist model have shown a promising start in dealing with sub-problems that are holistic in nature such as natural vs. man-made scenes, indoors vs. outdoors scenes and object prime location [25] , which can help to cue an appropriate direction. Salient objects help to create distinct signature of individual scenes that may not be differentiable by gist alone. They also provide a finer localization point of reference within each scene as well as between scenes. In the future we would like to present a physical implementation of this cooperative system, using bottom-up salient cues as well as context, to produce a useful topographical map for navigation in unconstrained world.

References

[1] Abe, Y., M. Shikano, T. Fukuda, F. Arai, and Y. Tanaka, "Vision Based Navigation System for Autonomous Mobile Robot with Global Matching," *IEEEInternational Conference on Robotics and Automation*, May 1999, 1299-1304.

[2] Hyvrinen A., "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, 10(3):626-634. 1999.

[3] Fox, D., W. Burgard, F. Dellaert, and S. Thrun, "Monte Carlo Localization: Efficient Position Estimation for Mobile Robots," *Proc. of the Sixteenth National Conference on Artificial Intelligence* (AAAI'99).

[4] Itti, L., "Models of Bottom-Up and Top-Down Visual Attention," *California Institute of Technology. Ph.D. thesis*, Jan 2000.

[5] Itti, L., "Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention," *IEEE Transactions on Image Processing*, in press, 2004.

[6] Katsura, H., J. Miura, M. Hild, and Y. Shirai "A View-Based Outdoor Navigation Using Object Recognition Robust to Changes of Weather and Seasons" *Intelligent Robots and Systems*, pp. 2974-2979, 2003.

[7] Leonard, J. J., and H. F. Durrant-Whyte, "Mobile robot localization by tracking geometric beacons," *IEEE Trans. Robotics and Automation*, 7(3):376-382, June 1991.

[8] Lingemann, K, H. Surmann, A. Nuchter, and J. Hertzberg, "Indoor and Outdoor Localization for Fast Mobile Robots" *IROS* 2004

[9] Maeyama, S., A. Ohya, and S. Yuta "Long distance outdoor navigation of an autonomous mobile robot by playback of Perceived Route Map" *ISER'97 Fifth International Symposium on Experimental Robotics*, pp.141-150, Jun. 1997.

[10] Matsumoto, Y., M. Inaba, H. Inoue, "View-Based Approach to Robot Navigation", *IEEE-IROS* , pp.1702-1708, 2000.

[11] Murrieta-Cid, R., C. Parra, M. Devy, "Visual Navigation in Natural Environments: From Range and Color Data to a Landmark-based Model," *Autonomous Robots* Vol. 13 no 2 pp. 143-168.

[12] Rensink, R.A., "The Dynamic Representation of Scenes," *Visual Cognition*, 7:17-42, 2000.

[13] Oliva, A., and P.G. Schyns, "Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli," *Cognitive Psychology*, 34, 72-107, 1997.

[14] Oliva, A., and P.G. Schyns, "Colored diagnostic blobs mediate scene recognition," *Cognitive Psychology*, 41, 176 - 210, 2000.

[15] Oliva, A., and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Intl J. Computer Vision*, vol.42, no.3, pp. 145-175, 2001.

[16] Potter, M. C., "Meaning in Visual Search," *Science*, pg.965-966, 187(4180), 1975.

[17] Potter, M. C., "Short-term conceptual memory for pictures," *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509-522 1976.

[18] Renniger, L.W., and J. Malik. "When is scene identification just texture recognition?," *Vision Research, 44 (2004)*, pp. 2301- 2311.

[19] Schyns, P. G., and A. Oliva. "From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition," *Psychological Science, 5(4)*, 195 - 200, 1994.

[20] Siagian, C., and L. Itti, "Biologically-Inspired Face Detection: Non-Brute-Force-Search Approach," *First IEEE International Workshop on Face Processing in Video*, June 2004.

[21] Takeuchi, Y., M. Hebert, "Evaluation of Image-Based Landmark Recognition Technique," *CMU-RI-TR-98-20*, Carnegie Mellon, 1998.

[22] Thorpe, S., D. Fize, and C. Marlot, "Speed of Processing in the Human Visual System," *Nature*, 381, 520 - 522, 1995.

[23] Thrun, S., D. Fox, and W. Burgard. "A probabilistic approach to concurrent mapping and localization for mobile robots," *Machine Learning*, 31:29-53, 1998. Also appeared in Autonomous Robots 5, 253-271.

[24] Thrun, S. "Finding Landmarks for Mobile Robot Navigation," *IEEE-ICRA*, pp. 958-963, May 1998.

[25] Torralba, A., and P. Sinha, "Statistical Context Priming for Object Detection," *IEEE Proc. Of Int. Conf in Comp. Vision*, 1:763-770, 2001.

[26] Torralba, A., K. P. Murphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition," *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France, October. 2003

[27] Tversky, B. and K. Hemenway, "Categories of The Environmental Scenes," *Cognitive Psychology*, 15, 121 - 149, 1983.

[28] Ulrich, I., and I. Nourbakhsh, "Appearance-based Place Rcognition for Topological Localization," *IEEE-ICRA*, pp. 1023 - 1029, April 2000.