Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention

Christian Siagian, *Member, IEEE,* Laurent Itti, *Member, IEEE,*

***Abstract*— We describe and validate a simple context-based scene recognition algorithm for mobile robotics applications. The system can differentiate outdoor scenes from various sites on a college campus using a multiscale set of early-visual features, which capture the "gist" of the scene into a low-dimensional signature vector. Distinct from previous approaches, the algorithm presents the advantage of being biologically plausible and of having low computational complexity, sharing its low-level features with a model for visual attention that may operate concurrently on a robot. We compare classification accuracy using scenes filmed at three outdoor sites on campus (13,965 to 34,711 frames per site). Dividing each site into nine segments, we obtain segment classification rates between 84.21% and 88.62%. Combining scenes from all sites (75,073 frames in total) yields 86.45% correct classification, demonstrating generalization and scalability of the approach.**

***Index Terms*— Gist of a scene, saliency, scene recognition, computational neuroscience, image classification, image statistics, robot vision, robot localization.**

I. Introduction

BUILDING the next generation of mobile robots hinges on solving tasks such as localization, mapping, and navigation. These tasks critically depend on developing capabilities to robustly answer the central question: Where are we? A significant number of mobile robotics approaches address this fundamental problem by utilizing sonar, laser, or other range sensors [1]–[3]. They are particularly effective indoors due to many spatial and structural regularities, including flat walls and narrow corridors. In the outdoors, however, these sensors become less robust because the structure of the enviroment can vary tremendously. It then becomes hard to predict the sensor input given all the protrusions and surface irregularities [4]. For example, a slight change in pose can result in large jumps in range reading because of tree trunks, moving branches, and leaves.

These difficulties with traditional robot sensors have prompted research towards other ways to obtain navigational input, especially by using the primary sensory modality of humans: vision. Within Computer Vision (there are several different approaches, listed below), lighting (especially in the outdoors), dynamic backgrounds, and view-invariant matching become major hurdles to overcome.

A. Object-Based Scene Recognition

A large portion of the vision-based approach towards scene recognition is object-based [5]–[7]. That is, a physical location is recognized by identifying a set of landmark objects (and possibly their configuration) known to be present at that location. This typically involves intermediate steps such as segmentation, feature grouping, and object recognition. Such layered approach is prone to carrying over and amplifying low-level errors along the stream of processing. For example, upstream identification of very small objects (pixel-wise) is hindered by downstream noise inherent to camera sensors, and by variable lighting conditions. This is particularly problematic in spacious environments like the open outdoors, where the landmarks tend to be more spread out and possibly at farther distances from the agent. It should also be pointed out that this approach needs to be environment-specific for the simplicity of selecting a small set of anchor objects, and that deciding on reliable and persistent candidate objects as landmarks is an open problem.

In recent years Scale Invariant Feature Transform (SIFT) [8] has been used in robotics quite extensively. We put SIFT in the object-based category because it is still a local feature and its usage are tied to the existance of reliable distinctive sub-structures, an object. This is especially true if the background

C. Siagian and L. Itti are with the University of Southern California, Departments of Computer Science, Psychology, and Neuroscience Program, Hedco Neuroscience Building - Room 30A, 3641 Watt Way, Los Angeles, California, 90089-2520. Correspondence should be addressed to itti@usc.edu.

is close to textureless (wide open spaces, the sky). On the other extreme, distracting background with too much texture that are mostly ephemeral (moving vegitation inside a forest) may also be too much to deal with, increasing the number of keypoints. On visually distinctive backgrounds, however, this method can be much more powerful than strictly object-based. Some systems [9], [10] bypass the segmentation stage and use the whole scene is one landmark. Because SIFT also allows for partial recognition, foreground objects, which tend to be less reliable (people walking, etc.), can be treated as distractions and the background become the prevailing strength (so long as the foreground object does not dominate the scene). And because SIFT's ability to absorb variability from out-of-plane rotation can only go so far, systems are also burdened with the need to store a large number of key-points from multiple views, which do not scale well.

B. Region-Based Scene Recognition

A different set of approaches [11]–[13] eliminates landmark objects, instead using segmented image regions and their configurational relationships to form a signature of a location. At this level of representation, the major problem is reliable region-based segmentation in which individual regions have to be robustly characterized and associated. Naïve template matching involving rigid relationships is often not flexible enough in the face of over/under-segmentation. This is especially true with unconstrained environments such as a park where vegetation dominates (refer to Experiment 2 in the present study). As a remedy, one can combine object-based and region-based approaches [13] by using regions as an intermediate step to locate landmark objects. Nevertheless, such technique is still prone to the same caveats as the object-based approach.

C. Context-Based Scene Recognition

The last set of approaches, which is context-based, bypasses the above traditional processing steps. Context-based approaches consider the input image as a whole and extract a low-dimensional signature that compactly summarizes the image's statistics and/or semantics. One motivation for such approach is that it should produce more robust solutions because random noise, which may catastrophically influence local processing, tends to average out globally. By identifying whole scenes, and not small sets of objects or precise region boundaries within the scenes, context-based approaches do not have to deal with noise and low-level image variations in small isolated regions, which plague both region segmentation and landmark recognition. The challenge to discover a compact and holistic representation for unconstrained images has hence prompted significant recent research. For example, Renniger and Malik [14] use a set of texture descriptors, and histogram to create an overall profile of an image. Ulrich and Nourbakhsh [15] build color histograms and perform matching using a voting procedure. In contrast, Oliva and Torralba [16] also encode some spatial information by performing 2D Fourier Transform analyses in individual image sub-regions on a regularly-spaced grid. The resulting spatially-arranged set of signatures, one per grid region, is then further reduced using principal component analysis (PCA) to yield a unique low-dimensional image classification key. Interestingly, Torralba reports that the entries in the key vector sometimes tend to correlate with semantically-relevant dimensions, such as city vs. nature, or beach vs. forest. In more recent implementations, Torralba [17] also used steerable wavelet pyramids instead of the Fourier transform.

D. Biologically-Plausible Scene Recognition

Despite all the recent advances in computer vision and robotics, humans still perform orders of magnitude better than the best available vision systems in outdoors localization and navigation. As such, it is inspiring to examine the low-level mechanisms as well as the system-level computational architecture according to which human vision is organized. Early on, the human visual processing system already makes decisions to focus attention and processing resources onto those small regions within the field of view which look more interesting. The mechanism by which very rapid holistic image analysis gives rise to a small set of candidate salient locations in a scene has recently been the subject of comprehensive research efforts and is fairly well understood [18]–[21].

Parallel with attention guidance and mechanisms for saliency computation, humans demonstrate exquisite ability at instantly capturing the "gist" of a scene; for example, following presentation of a photograph for just a fraction of a second, an observer may report that it is an indoor kitchen scene with numerous colorful objects on the countertop [22]–[25]. Such report at a first glance onto an image is remarkable considering that it summarizes the quintessential characteristics of an image, a process previously expected to require much analysis. With very brief exposures (100ms or below), reports are typically limited to a few general semantic attributes (e.g., indoors, outdoors, office, kitchen) and a coarse evaluation of distributions of visual features (e.g., highly colorful, grayscale, several large masses, many small objects) [26], [27]. However, answering specific questions such as whether an animal was present or not in the scene can be performed reliably down to exposure times of 28ms [28], [29], even when the subject's attention is simultaneously engaged by another concurrent visual discrimination task [30]. Gist may be computed in brain areas which have been shown to preferentially respond to "places," that is, visual scene types with a restricted spatial layout [31]. Spectral contents and color diagnosticity have been shown to influence gist perception [25], [32], leading to the development of the existing computational models that emphasize spectral analysis [33], [34].

In what follows, we use the term gist in a more specific sense than its broad psychological definition (what observers can gather from a scene over a single glance): we formalize gist as a relatively low-dimensional (compared to a raw image pixel array) scene representation which is acquired over very short time frames, and we thus represent gist as a vector in some feature space. Scene classification based on gist then becomes possible if and when the gist vector corresponding to a given image can be reliably classified as belonging to a given scene category.

From the point of view of desired results, gist and saliency appear to be complementary opposites: finding salient locations requires finding those image regions which stand out by significantly differing from their neighbors, while computing gist involves accumulating image statistics over the entire scene. Yet, despite these differences, there is only one visual cortex in the primate brain, which must serve both saliency and gist computations. Part of our contribution is to make the connection between these two crucial components of biological vision. To this end, to be biologically-plausible, we here explicitly explore whether it is possible to devise a working system where the low-level feature extraction mechanisms — coarsely corresponding to cortical visual areas V1 through V4 and MT — are shared and serve both attention and gist, as opposed to computed separately by two different machine vision modules.

The divergence comes at a later stage, in how the low-level vision features are further processed before being utilized. In our neural simulation of posterior parietal cortex along the dorsal or "where" stream of visual processing [35], a saliency map is built through spatial competition of low-level feature responses throughout the visual field. This competition quiets down locations which may initially yield strong local feature responses but resemble their neighbors, while amplifying locations which have distinctive appearances. In contrast, in our neural simulation of inferior temporal or the "what" stream of visual processing, responses from the low-level feature detectors are combined to produce the gist vector as a holistic low-dimensional signature of the entire input image. The two models, when run in parallel, can help each other and provide a more complete description of the scene in question. Figure 1 shows a diagram of our implementation.

In the present paper, our focus is on image classification using the gist signature computed by this model, while exploitation of the saliency map has been extensively described previously for a number of vision tasks [20], [21], [36], [37]. We describe, in the following sections, our algorithm to compute gist in a very inexpensive manner by using the same low-level visual front-end as the saliency model. We then extensively test the model in three challenging outdoor environments across multiple days and times of days, where the dominating shadows, vegetation, and other ephemerous phenomena are expected to defeat landmark-based and region-based approaches. Our success in achieving reliable performance in each environment is further generalized by showing that performance does not degrade when combining all three environments. These results support our hypothesis that gist can reliably be extracted at very low computational cost, using very simple visual features shared with an attention system in an overall biologically-plausible framework.

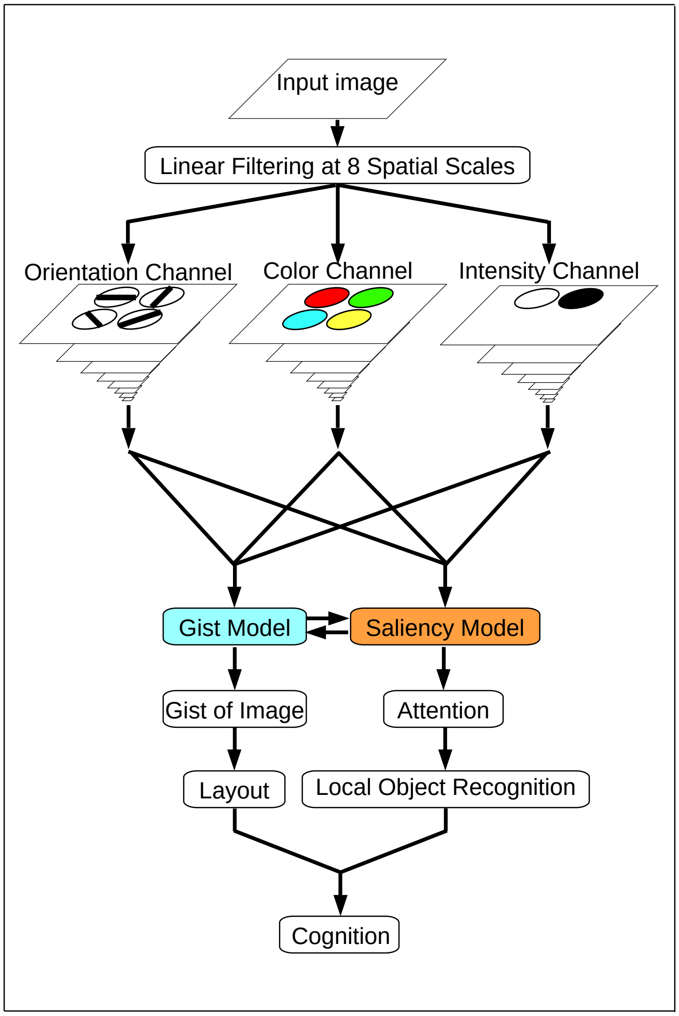


Fig. 1. Model of Human Vision with Gist and Saliency

II. Design and Implementation

The core of our present research focuses on the process of extracting the gist of an image using features from several domains, calculating its holistic characteristics but still taking into account coarse spatial information. The starting point for the proposed new model is the existing saliency model of Itti *et al.* [20], [38], which is freely available on the World-Wide-Web.

A. Visual Cortex Feature Extraction

In the saliency model, an input image is filtered in a number of low-level visual "feature channels" at multiple spatial scales, for features of color, intensity, orientation, flicker and motion (found in Visual Cortex). Some channels (color, orientation, and motion) have several sub-channels (color type, orientation, direction of motion). Each sub-channel has a nine-scale pyramidal representation of filter outputs, a ratio of 1:1 (level 0) to 1:256 (level 8) in both horizontal and vertical dimensions, with a 5-by-5 Gaussian smoothing applied in between scales. Within each sub-channel i, the model performs center-surround operations (commonly found in biological-vision which compares image values in center-location to its neighboring surround-locations) between filter output maps, $O_i(s)$, at different scales s in the pyramid. This yields feature maps $M_i(c, s)$, given "center" (finer) scale c and "surround" (coarser) scale s. Our implementation uses $c = 2, 3, 4$ and $s = c+d$, with $d = 3, 4$. Across-scale difference (operator $\ominus$) between two maps is obtained by interpolation to the center (finer) scale and pointwise absolute difference (eqn. 1).

For color and intensity channels:

$$M_i(c, s) = |O_i(c) \ominus O_i(s)| = |O_i(c) - \text{Interp}_{s-c}(O_i(s))| \tag{1}$$

Hence, we compute six feature maps for each type of feature at scales 2-5, 2-6, 3-6, 3-7, 4-7, and 4-8, so that the system can gather information in regions at several scales, with added lighting invariance provided by the center-surround comparison (further discussed below).

In the saliency model, feature maps were used to detect conspicuous regions in each channel, through additional winner-take-all mechanisms which emphasize locations which substantially differ from their neighbors [20]. The feature maps are then linearly combined to yield a saliency map.

To re-use the same low-level maps for gist as for attention, our gist model uses the already available orientation, color and intensity channels. Flicker and motion channels, which also contribute to the saliency map, are assumed to be more dominantly determined by the robot's egomotion and hence unreliable in forming a gist signature of a given location. Our basic approach is to exploit statistical data of color and texture measurements in predetermined region subdivisions. These features are independent of shape as they are simply denoting lines and blobs. We incorporate information from the orientation channel, employing Gabor filters to the greyscale input image (eqn. 2) at four different angles ($\theta_i = 0, 45, 90, 135°$) and at four spatial scales ($c = 0, 1, 2, 3$) for a subtotal of sixteen sub-channels.

For orientation channels:

$$M_i(c) = \text{Gabor}(\theta_i, c) \quad (2)$$

We do not perform center-surround on the Gabor filter outputs because these filters already are differential by nature. The color and intensity channels combine to compose three pairs of color opponents derived from Ewald Hering's Color Opponency theories [39], which identify four primary colors red, green, blue, yellow (denoted as R,G,B, and Y in eqns. 3, 4, 5, and 6, respectively) and two hueless dark and bright colors 7, computed from the raw camera r, g, b outputs [20].

$$R = r - (g + b)/2 \quad (3)$$

$$G = g - (r + b)/2 \quad (4)$$

$$B = b - (r + g)/2 \quad (5)$$

$$Y = r + g - 2(|r - g| + b) \quad (6)$$

$$I = (r + g + b)/3 \quad (7)$$

The color opponency pairs are the two color channels' red-green and blue-yellow (eqn. 8 and 9), along with the intensity channel's dark-bright opponency (eqn. 10). Each of the opponent pairs are used to construct six center-surround scale combinations. These eighteen sub-channels along with the sixteen Gabor combinations add up to a total of thirty-four sub-channels altogether. Because the present gist model is not specific to any domain, other channels such as stereo could be used as well.

$$RG(c,s) = |(R(c) - G(c)) \ominus (R(s) - G(s))| \quad (8)$$

$$BY(c,s) = |(B(c) - Y(c)) \ominus (B(s) - Y(s))| \quad (9)$$

$$I(c,s) = |I(c) \ominus I(s)| \quad (10)$$

Figure 2 illustrates the gist model architecture.

B. Gist Feature Extraction

After the low-level center-surround features are computed, each sub-channel extracts a gist vector from its corresponding feature map. We apply averaging operations (the simplest neurally-plausible computation) in a fixed four-by-four grid of subregions over the map. Observe figure 3 for visualization of the process.

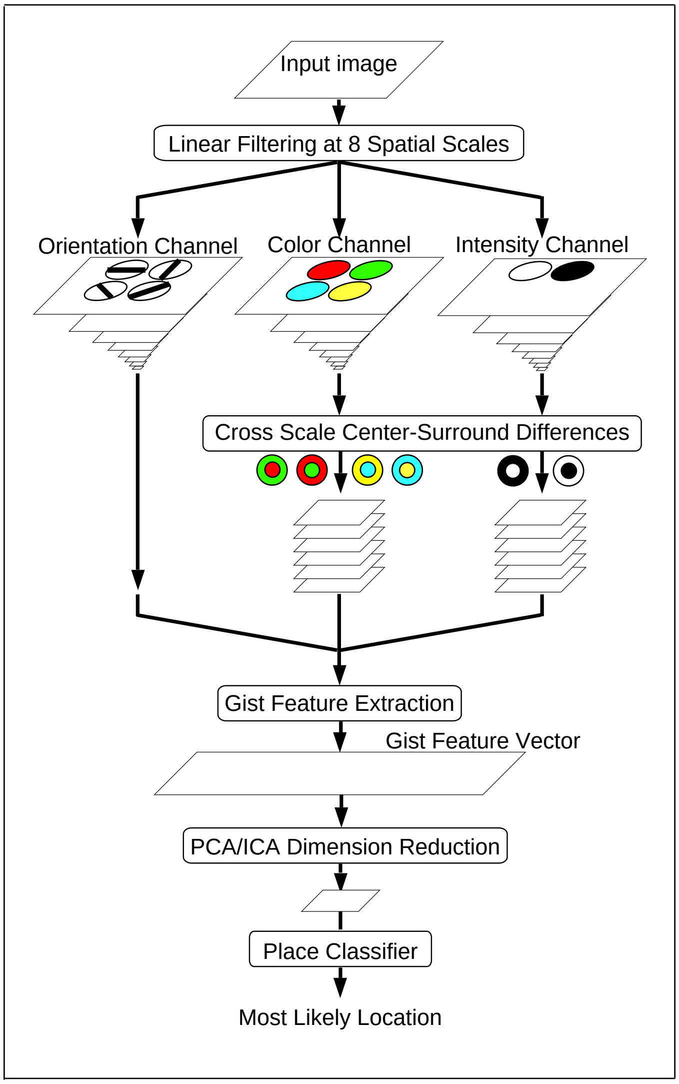


Fig. 2. Visual Feature Channels Used in the Gist Model

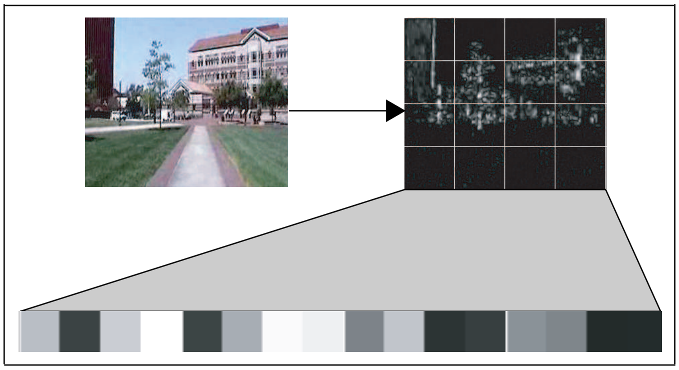


Fig. 3. Gist Decomposition of Vertical Orientation Sub-channel. The original image (top left) is put through a vertically-oriented Gabor filter to produce a feature map (top right). That map is then divided into 4-by-4 grid sub-regions. We then take the mean of each grid to produce 16 values for then gist feature vector (bottom).

Thus, as proposed in introduction, gist cumulates information over space in image sub-regions, while saliency relied on competition across space. Equation 11 formalizes the computation for each of the sixteen raw gist features ($G_i^{k,l}(c,s)$) per map, taking the sum, over a given sub-region (specified by indices k and l in the horizontal and vertical direction, respectively), of the values in $M_i(c,s)$, then dividing by the number of pixels in the sub-region.

For color, intensity channels:

$$G_i^{k,l}(c,s) = \frac{1}{16WH} \sum_{u=\frac{kW}{4}}^{\frac{(k+1)W}{4}-1} \sum_{v=\frac{lH}{4}}^{\frac{(l+1)H}{4}-1} [M_i(c,s)](u,v) \tag{11}$$

Where W and H are the width and height of the entire image. We similarly process the orientation maps $M_i(c)$ to compute $G_i^{k,l}(c)$.

Although additional statistics such as variance would certainly provide useful additional descriptors, their computational cost is much higher than that of first-order statistics and their biological plausibility remains debated [40]. Thus, here we explore whether first-order statistics would be sufficient to yield reliable classification, if one relies on using the available variety of visual features to compensate for more complex statistics within each feature.

C. Color Constancy

The advantage of coarse statistical-based gist is stability in averaging out local and random noise. What is more concerning is global bias such as lighting as it changes the appearance of the entire image. Color constancy algorithms such as gray world and white patch assume that lighting is constant throughout a scene [41], [42]. Unfortunately, outdoor ambient light is not quite as straightforward. Not only does it change over time, both in luminance and chrominance, but also vary within a single scene as it is not a point light-source. Different sun positions and atmospheric conditions illuminate different parts of a scene in varying degrees as illustrated by images taken one hour apart, juxtaposed in the first row of figure 4. We can see that the foreground of image 1 receives more light while the background does not. Conversely, the opposite occurs in image 2. It is important to note that the goal of the step is not to recognize/normalize color with high accuracy, but to produce stable gist features over color and intensity. We also consider another (iterative and slower converging) normalization technique called Comprehensive Color Normalization (CCN) [43], which can be seen as both global and local.

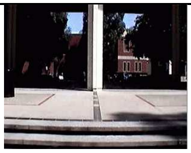
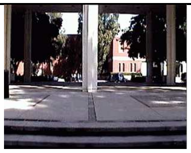
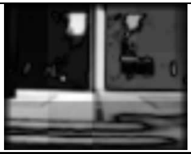
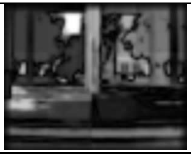
Channel	Image 1	Image 2	Sub-channel	pSNR (db)
Raw			r g b	9.24 9.60 10.08
Green Red - Green			2 & 5 2 & 6 3 & 6 3 & 7 4 & 7 4 & 8	32.57 32.13 34.28 33.95 36.32 35.82
Blue - Yellow			2 & 5 2 & 6 3 & 6 3 & 7 4 & 7 4 & 8	32.44 30.83 32.42 30.95 31.95 31.95
Dark- Bright			2 & 5 2 & 6 3 & 6 3 & 7 4 & 7 4 & 8	15.03 12.72 13.79 12.21 13.29 13.33

Fig. 4. Example of two lighting conditions of the same scene. pSNR (peak signal-to-noise ratio) values measure how similar the maps are between images 1 and 2. Higher pSNR values for a given map indicate better robustness of that feature to variations in lighting conditions. Our center-surround channels exhibit better invariance than the raw r, g, b channels although, obviously, are not completely invariant.

One indisputable fact is that when texture is lost because of lighting saturation (both too bright or too dark for the camera sensor), no normalization, however sophisticated, can bring it back. To this end, because of the nature of our gist computation, the best way is to recognize gists of scenes with different lighting separately. We thus opted not to add any preprocessing, but instead to train our gist classifier (described below) on several lighting conditions. The gist features themselves already helped minimize the effect of illumination change because of their differential nature (Gabor or center-surround). Peak Signal-to-Ratio (pSNR) tests for the two images with differing lighting conditions in figure 4 show better invariance for our differential features than for the raw r, g, b features, especially for the two opponent color channels. This shows that the low-level feature processing produces contrast information that is reasonably robust to lighting. Note that using differential features comes at a price: baseline information (e.g., absolute color distributions) is omitted from our gist encoding even though it has been shown to be useful [15]. The center-surround can be construed as only looking for edges surrounding regions (and not the regions themselves), because that is where the contrasts are. On the other hand, calculating color distribution histograms amounts to measuring the size of those regions. This is where the pyramid scheme helps recover some of the information. With the pyramid scheme the system can pick up regions, at coarser scales [25] and indirectly infer the absolute distribution information with the added lighting invariance. As an example, the intensity channel output for the illustration image of figure 5 shows different-sized regions being emphasized according to their respective center-surround parameter.

D. PCA/ICA Dimension Reduction

The total number of raw gist feature dimension is 544, 34 feature maps times 16 regions per map (figure 5). We reduce the dimensions using Principal Component Analysis (PCA) and then Independent Component Analysis (ICA) with FastICA [44] to a more practical number of 80 while still preserving up to 97% of the variance for a set in the upwards of 30,000 campus scenes.

E. Scene Classification

For scene classification, we use a three-layer neural network (with intermediate layers of 200 and 100 nodes), trained with the back-propagation algorithm done on a 1.667GHz Athlon AMD machine. The main reason for using a neural network classifier is the success rate (see Results) which proves that it is adequate. In addition, it is easy to add more samples and the training process takes a short amount of time. The complete process is illustrated in figure 5.

III. Testing and Results

We test the system at several sites on campus (map shown in figure 6). The first one is the Ahmanson Center for Biological Research (ACB), in which the scenes are filmed around the building complex. Most of the surroundings are flat walls with little texture and solid lines that delineate the walls and different parts of the buildings. A region-based representation would find the environment ideal. Figure 7 shows some of the scenes around ACB with their corresponding visual depiction of gist. The second site is a region comprised of two adjoining parks: Associate and Founders park (AnF) which are dominated by vegetations. Large areas of the images are practically un-segmentable as leaves overrun most regions. And although there are objects such as lamp posts and benches, lighting inside the park may often render their recognition difficult because the sunlight is randomly blocked by the trees. Refer to Figure 8 for the variety of the visual stimuli collected along the path. The third site we tested are an open area in the Frederick. D. Fagg park. A large portion of the scenes (figure 9) is the sky, mostly textureless space with random light clouds.

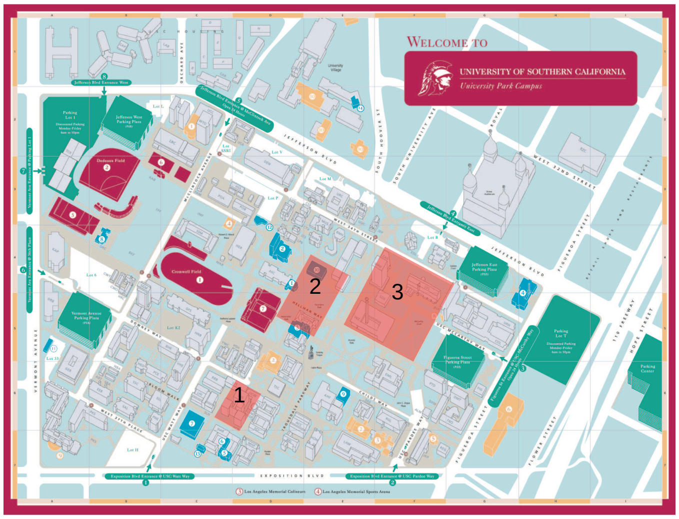


Fig. 6. Map of the experiment sites

To collect visual data we use an 8mm handheld camcorder. The captured video clips are hardly stable as the camera is carried by a person walking while filming. Moreover, attempts to smooth out image jitters are not always successful. At this point the data is still view-specific, as each location is only traversed from one direction. For a view-invariant scene recognition, we need to train the system on multiple views [17]. We have tried to sweep the camera left to right (and vice versa) to create a wider point of view, although to retain performance the sweep has to be done at a much slower pace than regular walking speed [45]. The current testing setup is slightly selective as the amount of foreground interference is minimized by filming during off-peak hours where fewer people

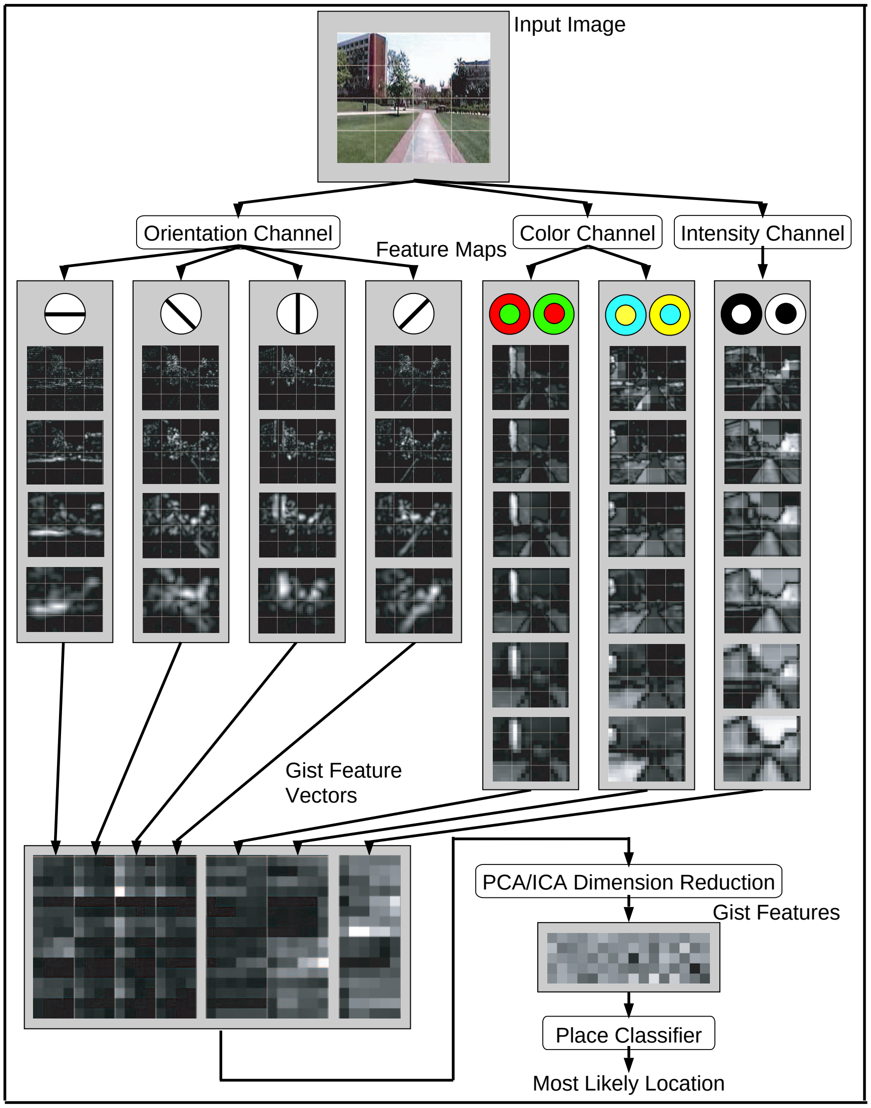


Fig. 5. Example of Gist Feature Extraction

are out walking. It should be pointed out that gist can include foreground objects as part of a scene as long as they do not dominate it.

For classification, we divide the video clips into segments. A segment is a portion of a hallway, path, or road interrupted by a crossing or a physical barrier at both ends. The path is divided this way because, along with natural geographical delineation, images within each segment look similar to a human observer. Moreover, when separating the data at the junction, we take special care in creating a clean break between the two involved segments. That is, we stop short of a crossing for the current segment and wait a few moments before starting the next one. This ansures that the system will be trained with data where ground-truth labeling (assigning a segment number to an image) is unambiguous. In addition, for the frames in the middle of the clips we include all of them, there is no by-hand selection being done.

The main issue in collecting training samples is the selection of filming times that include all lighting conditions. Because lighting space is hard to gauge, we perform trial-and-error to come up with times of the day which attempt to cover the space. We take data on up to six different times of the day, twice for each, for several days. They cover the brightest (noon time) to the darkest (early evening) lighting conditions, overcast vs. clear, and encompass notable changes in apperance due to increases in temperature (hazy mid-afternoon). Although we are bound to exclude some lighting conditions, the

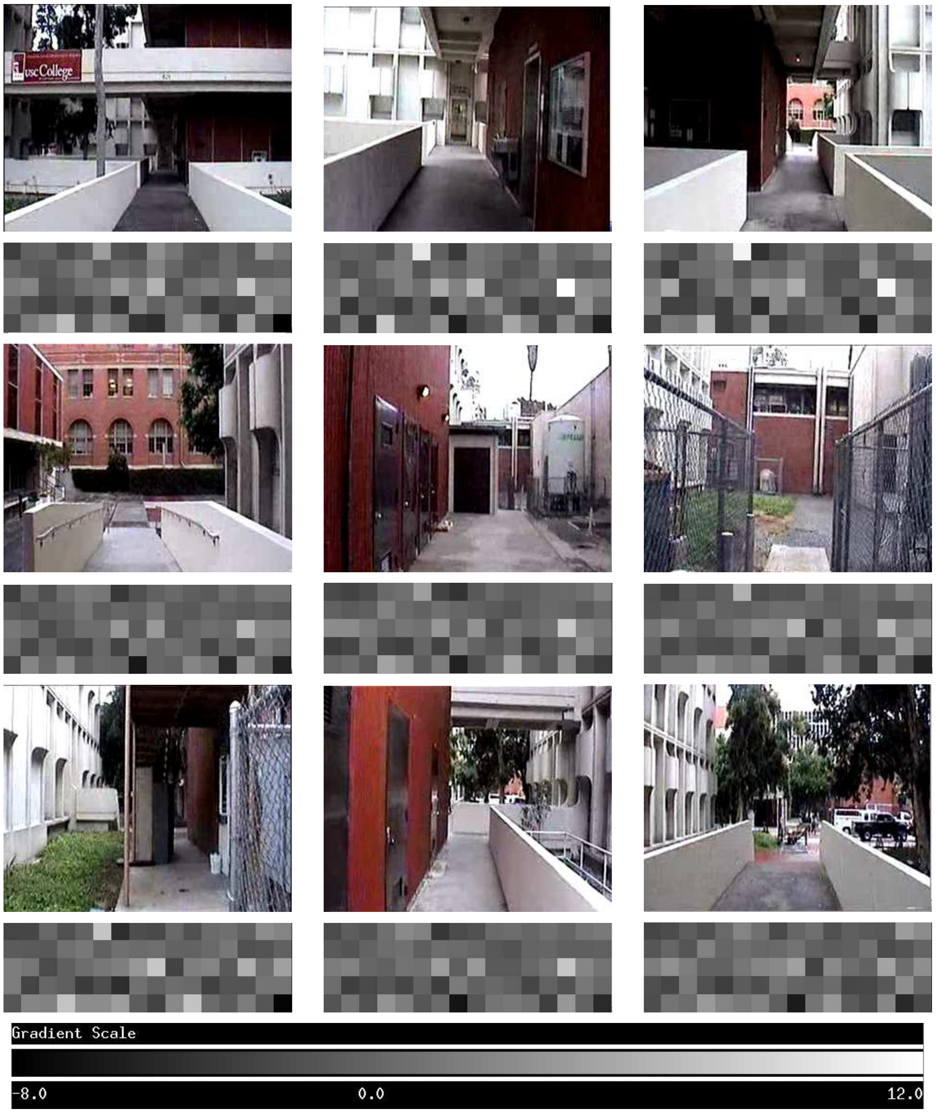


Fig. 7. Examples of images in each segment of ACB

results show that the collected samples cover a large portion of the space. It should be noted that most (10 of 12) of the training and testing data are taken on different days. For the two taken on the same day, the testing data was recorded is in the early evening (dark lighting) while training data was recorded near noon (bright lighting).

Each of the first three experiments uses the same classifier neural network structure, with nine output layer nodes (same as the number of segments). We use absolute encoding for the training data. For example, if the correct answer for an image is segment 1, the corresponding node is assigned 1.0, while the others are all 0.0. The encoding allows for probabilistic outputs for scenes in the testing data. For completeness, the intermediate layers have 200 and 100 nodes, respectively, while we have 80 input nodes (for the 80 features of the gist vector). That is a total of: 80*200 + 200*100 + 100*9 = 36,900 connections. Our cut-off for convergence is 1% training error. All training is done on a 1.667GHz Athlon AMD machine.

A. Experiment 1: Ahmanson Center for Biological Research (ACB)

This experimental site was chosen to investigate what the system can achieve in a rigid and less spacious man-made environment. Each segment is a straight line and part of a hallway. Some hallways are divided into two segments so that each segment is approximately of the same length. Figure 10 shows the map of the segments while figure 7 displays scenes of each segment. Note that the

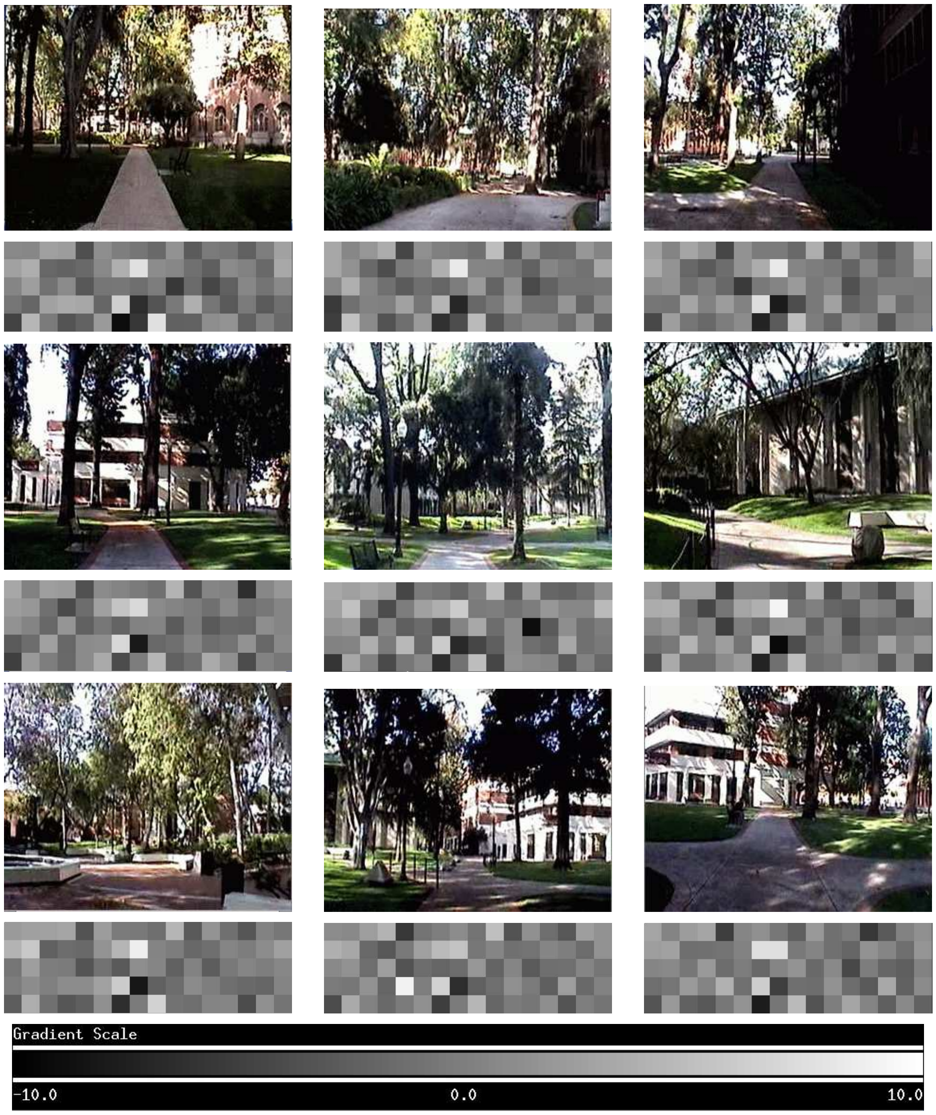


Fig. 8. Examples of images in each segment of Associate and Founders park

map shows that the segments are not part of a single continuous path, but a series of available walkways within an environment, traversed from a single direction.

Figure 11 represents the four lighting conditions used in testing: late afternoon, early evening (note that the lights are already turned on), noon, and mid-afternoon. We chose two dark and two bright conditions to assure a wide range of testing conditions.

We train the system several times before choosing a training run that gives the highest classification result on the training data. After only about twenty epochs, the network converges to less than 1% error. A fast rate of training convergence in the first few epochs appears to be a telling sign of how successful classification will be during testing. Table I shows results. The term “False+” for segment x means the number of incorrect segment x guesses given that the correct answer is another segment, divided by the total number frames in the segment; conversely, “False-” is the number of incorrect guesses given that the correct answer is segment x, divided the total number frames in the segment. The table shows that the system is able to classify the segments consistently during the testing phase with a total error of 12.04% or an overall 87.96% correctness.

We also report the confusion matrix in table II and find that the errors are, in general, not uniformly distributed. Spikes of classification errors between segments 1 and 2 would suggest a possibility of significant overlapping scenes (segment 2 is a continuation of segment 1; see figure 10). On the other hand, there are also errors that are not as easily explainable. For example, 163 false positives for

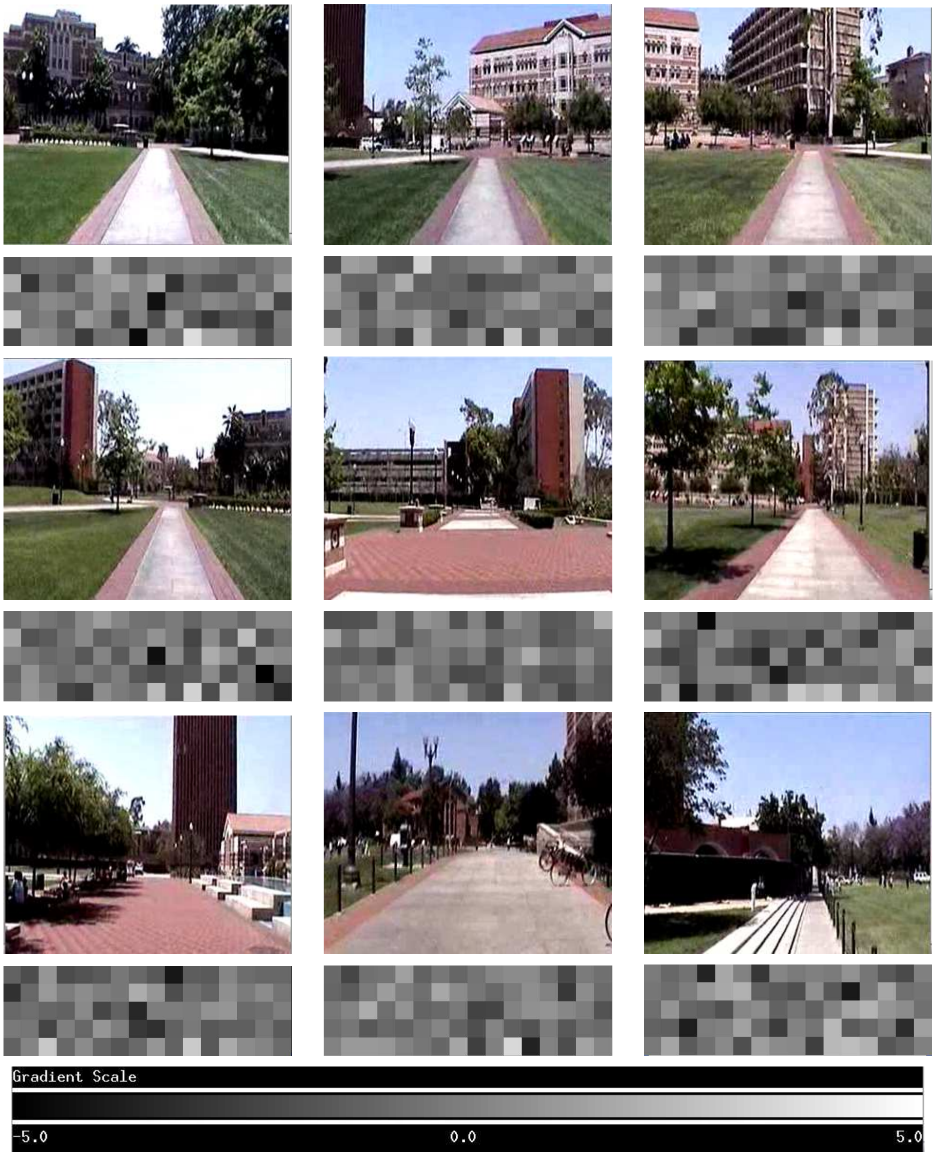


Fig. 9. Examples of images, one from each segment of Frederick D. Fagg park (segments 1 through 9 from left to right and top to bottom).

TABLE I

AHMANSON CENTER FOR BIOLOGY EXPERIMENTAL RESULTS

	Trial 1		Trial 2		Trial 3		Trial 4		Total			
Segment	False+	False-	False+	False-	False+	False-	False+	False-	False+	Percent.	False-	Percent.
1	14/390	11/387	17/380	47/410	32/393	27/388	39/445	5/411	102/1608	6.34%	90/1596	5.64%
2	20/346	114/440	133/468	101/436	85/492	54/461	18/325	131/438	256/1631	15.70%	400/1775	22.54%
3	1/463	3/465	0/456	29/485	82/502	43/463	33/475	31/473	116/1896	6.12%	106/1886	5.62%
4	7/348	18/359	24/338	7/321	5/226	84/305	7/148	108/249	43/1060	4.06%	217/1234	17.59%
5	46/348	5/307	52/389	0/337	64/290	95/321	125/403	41/319	287/1430	20.07%	141/1284	10.98%
6	24/567	13/556	39/478	56/495	23/533	24/534	69/564	7/502	155/2142	7.24%	100/2087	4.79%
7	43/410	71/438	55/371	129/445	136/439	95/398	108/486	22/400	342/1706	20.05%	317/1681	18.86%
8	101/391	0/290	18/265	0/247	67/320	21/274	37/303	22/288	223/1279	17.44%	43/1099	3.91%
9	65/320	86/341	46/404	15/373	17/262	68/313	29/227	98/296	157/1213	12.94%	267/1323	20.18%
Total	321/3583		384/3549		511/3457		465/3376		1681/13965			
Percent.	8.96%		10.82%		14.78%		13.77%		12.04%			

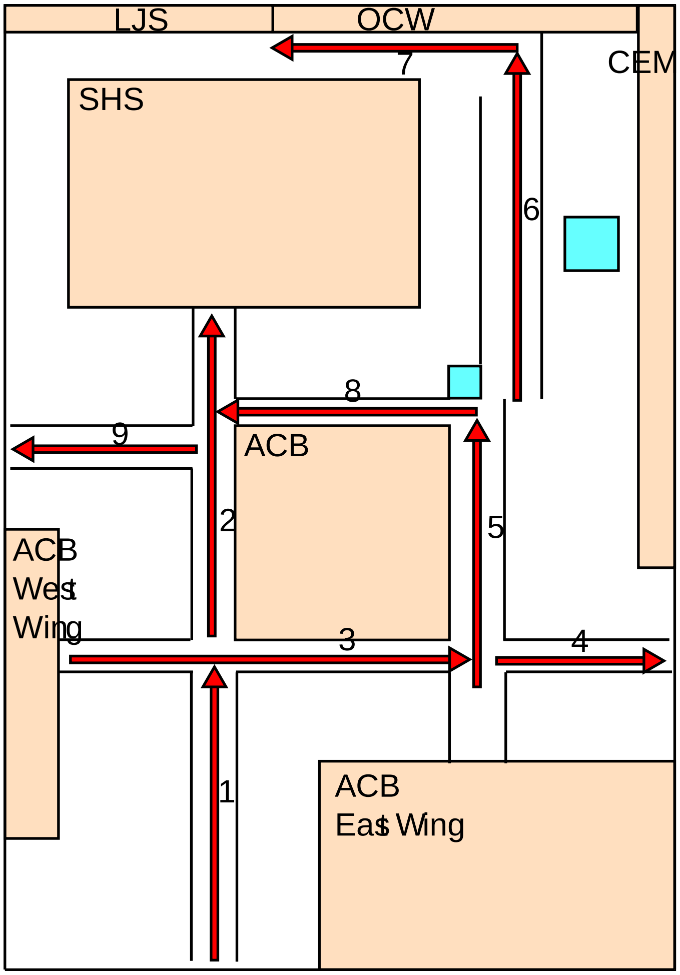


Fig. 10. Map of the path segments of ACB

segment 2 when the ground truth is segment 7 (and 141 false positives in the other direction). From figure 7 we can see that there is little resemblance between the two apperance-wise. However, if we consider just the coarse layout, the structure of both segments are similar, with a white region on the left side and a dark red one on the right side.

B. Experiment 2: Associates and Founders Park (AnF)

We now compare the results of Experiment 1 with, conceivably, a more difficult classification task: segmenting paths in a vegetation-dominated site. Figure III maps out the segments while figure 8 displays a sample image from each of them. As we can see there are fewer extractable structures. In addition, the lengths of the segments at this site are about twice the lengths of the segments in Experiment 1. As with Experiment 1, we perform multi-layer neural network classification using the back-propagation algorithm with the same network architecture and parameters. The number of epochs for training convergence is less than 40, about twice that found for Experiment 1.

Figure 13 shows four lighting conditions tested: early evening (lights already turned on), overcast, noon, and mid-afternoon. Also note that at the first test run, the bench in the front is missing from the image. We encounter similar challenges in other segments, such as service vehicles parked or a huge storage box placed in the park for a day.

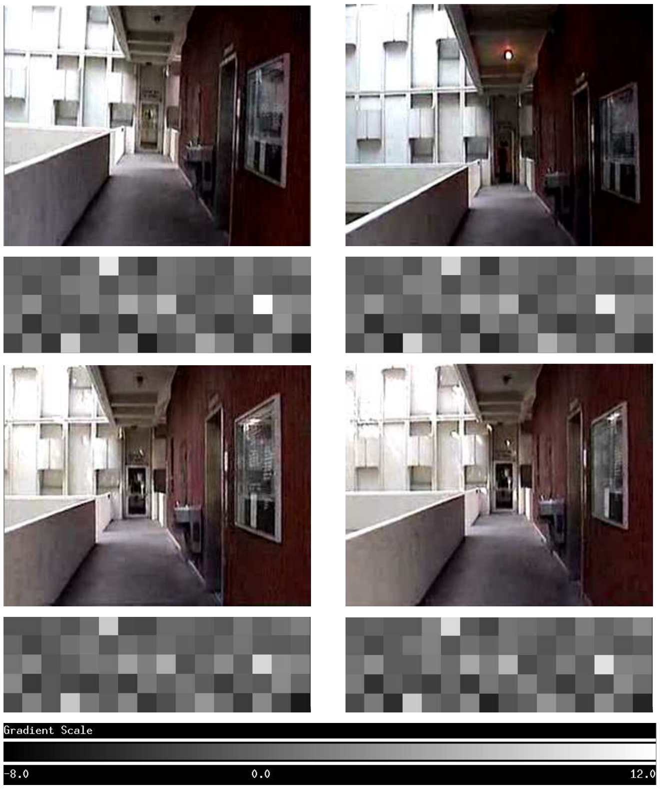


Fig. 11. Lighting conditions used for testing at Ahmanson Center for Biology (ACB). Clockwise from top left: late afternoon, early evening, noon, and mid-afternoon

TABLE II

AHMANSON CENTER FOR BIOLOGY CONFUSION MATRIX

Segment number guessed by algorithm

True segment number

Segment	1	2	3	4	5	6	7	8	9
1	1506	39	0	1	0	12	25	0	13
2	77	1375	0	0	0	54	141	11	117
3	1	6	1780	19	40	11	4	24	1
4	3	1	66	1017	66	49	14	18	0
5	0	10	4	0	1143	4	114	4	5
6	0	9	7	3	61	1987	10	10	0
7	18	163	0	13	1	19	1364	82	21
8	0	14	1	3	15	3	7	1056	0
9	3	14	38	4	104	3	27	74	1056

TABLE III
ASSOCIATE AND FOUNDERS PARK EXPERIMENTAL RESULTS

	Trial 1		Trial 2		Trial 3		Trial 4		Total			
Segment	False+	False-	False+	False-	False+	False-	False+	False-	False+	Percent.	False-	Percent.
1	71/559	210/698	177/539	440/802	140/786	245/891	49/733	62/746	437/2617	16.70%	957/3137	30.51%
2	38/544	64/570	107/429	6/328	271/558	187/474	122/584	12/474	538/2115	25.44%	269/1846	14.57%
3	57/851	71/865	54/814	217/977	206/1096	78/968	38/996	5/963	355/3757	9.45%	371/3773	9.83%
4	61/518	31/488	72/611	58/597	221/730	179/688	131/652	111/632	485/2511	19.32%	379/2405	15.76%
5	82/669	30/617	142/867	45/770	121/785	110/774	54/744	87/777	399/3065	13.02%	272/2938	9.26%
6	300/1254	47/1001	265/1210	177/1122	273/1084	192/1003	148/1079	167/1098	986/4627	21.31%	583/4224	13.80%
7	42/297	167/422	177/643	104/570	54/553	62/561	76/416	59/399	349/1909	18.28%	392/1952	20.08%
8	54/577	75/598	73/696	69/692	59/771	85/797	60/770	58/768	246/2814	8.74%	287/2855	10.05%
9	106/737	116/747	53/858	4/809	146/655	353/862	69/732	186/849	374/2982	12.54%	659/3267	20.17%
Total	811/6006		1120/6667		1491/7018		747/6706		4169/26397			
Percent.	13.50%		16.80%		21.25%		11.14%		15.79%			

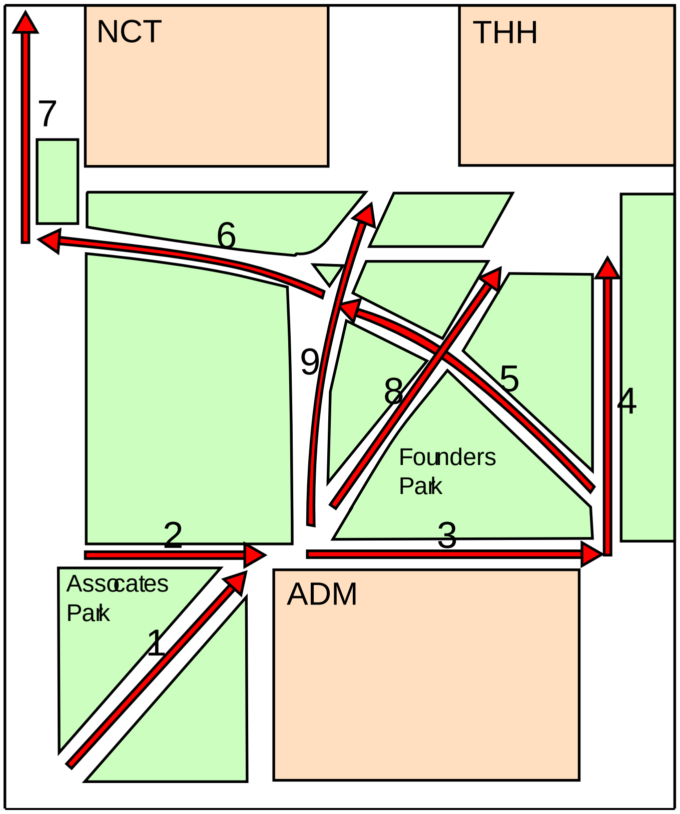


Fig. 12. Map of the path segments of Associate and Founders Park

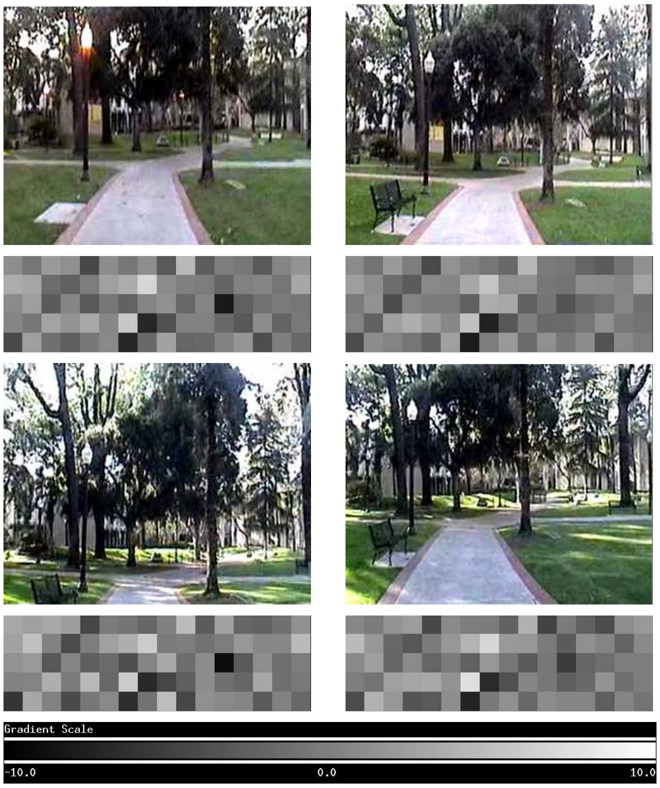


Fig. 13. Lighting conditions used for testing at Asociate and Founders park (AnF). Clockwise from top left: early evening, overcast, noon, and mid-afternoon

The result are shown at table III. The confusion matrix (table IV) is also reported. A quick glance at table III reveals that the performance, with a total error of 15.79% (84.21% success rate), is higher than in Experiment 1. However, if we look at the challenges presented by the scenes, it is quite an accomplishment to lose less than 4% in performance. In addition, no calibration is done in moving from the first environment to the second. Increases in length of segments do not affect the results drastically. The results from the third experiment which has even longer segments will confirm this assessment. It appears that the longer length does not mean more variability to absorb because the majority of the scenes within a segment do not change all that much. The confusion matrix (table IV) shows that the errors are marginally more

TABLE IV
ASSOCIATE AND FOUNDERS PARK CONFUSION MATRIX

		Segment number guessed by algorithm								
	Segment	1	2	3	4	5	6	7	8	9
True segment number	1	2180	32	21	28	212	374	96	52	142
	2	20	1577	43	7	1	186	7	4	1
	3	118	26	3402	73	51	4	27	33	39
	4	3	131	89	2026	4	66	45	5	36
	5	68	2	9	13	2666	49	8	105	18
	6	22	86	59	142	40	3641	161	6	67
	7	38	62	1	14	15	201	1560	4	57
	8	78	24	73	52	14	29	3	2568	14
	9	90	175	60	156	62	77	2	37	2608

uniform than in Experiment 1 (few zero entries), probably as the environment is less structured and prone to more accidental classification errors among possibly non-adjacent segments when vegetation dominates.

C. *Experiment 3: Frederick D. Fagg park (FDF)*

The third and final site is an open area in front of the Leavey and Doheny libraries called the Frederick D. Fagg park, which the students use to study outdoors and to catch some sun. The main motivation for testing at this site is to assess the gist response on sparser data. Figure 14 shows the map of segments while figure 9 shows the scenes from each segment. The segments are about 50% longer than the ones in the second experiment (three times that of experiment 1). The number of epochs in training goes up by about ten and the amount time of convergence roughly doubles from that of experiment 2, to about 50 minutes. Figure 15 represents the four lighting conditions tested: early evening (the street lights not yet turned on), evening (the street lights are already turned on), noon, and middle of afternoon.

Table V shows the results for the experiment, listing total error of 11.38% (88.62% classification). The result from trail 2 (7.95% error) is the best among all runs for all experiments. We suspect that this success is because the lighting very closely resembles the one of the training data. The run is conducted at noon, in which the lighting does tend to stay the same for long periods of time. As a performance reference, when we test the system with a set of data taken right after a training set, the error rates are about 9% to 11%. When training images with the same lighting condition as on a subset of testing data of interest are excluded during training, the error for that run usually at least triples (to about thirty to forty percent), which suggests that lighting coverage in the training phase is a critical factor.

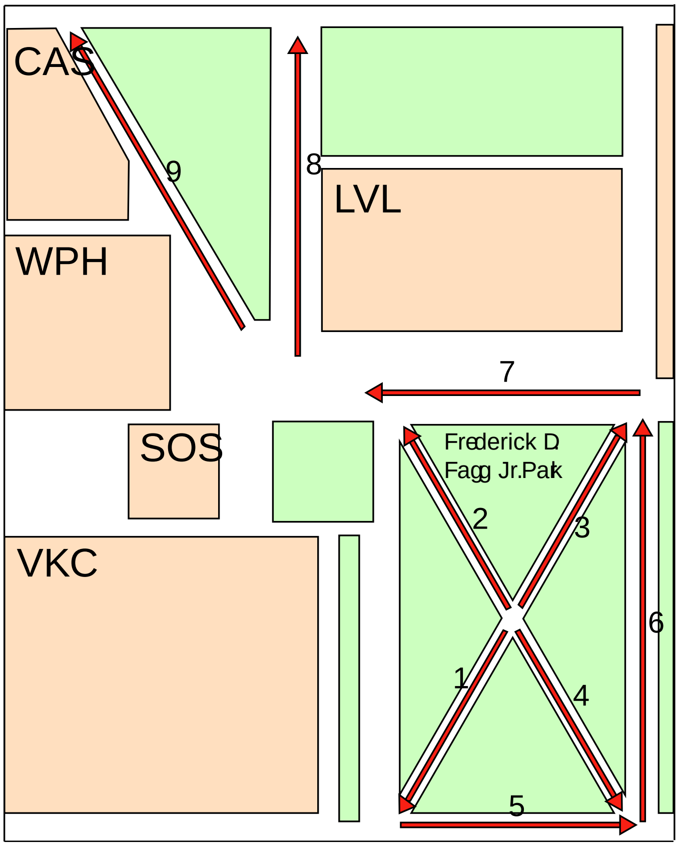


Fig. 14. Map of the path segments of Frederick D. Fagg park

The confusion matrix for Experiment 3 (table VI is also reported. Overall, the results are better than Experiment 1 and 2 even though the segments are longer on average. It can be argued that the system performance degrades gracefully with the subjectively-assessed visual difficulty of the environment, experiment 2 (AnF) being the most challenging one.

D. *Experiment 4: Combined sites*

To gauge the system's scalability, we combine scenes from all three sites and train it to classify twenty seven different segments. The only difference in the neural-network classifier is that the output layer now consists of twenty-seven nodes. The number of the input and hidden nodes remains the same. The number of connections is

TABLE V
FREDERICK D. FAGG PARK EXPERIMENTAL RESULTS

	Trial 1		Trial 2		Trial 3		Trial 4		Total			
Segment	False+	False-	False+	False-	False+	False-	False+	False-	False+	Percent.	False-	Percent.
1	22/657	246/881	40/699	11/670	44/684	207/847	28/735	246/953	134/2775	4.83%	710/3351	21.19%
2	246/1022	12/788	53/749	44/740	56/727	126/797	105/758	225/878	460/3256	14.13%	407/3203	12.71%
3	11/691	178/858	0/689	7/696	341/1218	45/922	282/1147	5/870	634/3475	16.93%	235/3346	7.02%
4	5/799	43/837	3/663	80/740	53/883	7/837	35/757	99/821	96/3102	3.09%	229/3235	7.08%
5	18/440	409/831	11/390	369/748	2/696	0/694	16/870	0/854	47/2396	1.96%	778/3127	24.88%
6	343/1976	47/1680	12/1550	27/1565	182/1772	122/1712	243/1770	145/1672	780/7068	11.04%	341/6629	5.14%
7	0/806	231/1037	25/944	4/923	0/675	182/857	30/886	38/894	55/3311	1.66%	455/3711	12.26%
8	483/1607	48/1172	436/1568	79/1211	319/1581	93/1355	149/1244	175/1270	1387/6000	23.12%	395/5008	7.89%
9	86/825	0/739	65/866	24/825	42/579	257/794	164/788	119/743	357/3058	11.67%	400/3101	12.90%
Total	1214/8823		645/8118		1039/8815		1052/8955		3950/34711			
Percent.	13.76%		7.95%		11.787%		11.75%		11.38%			

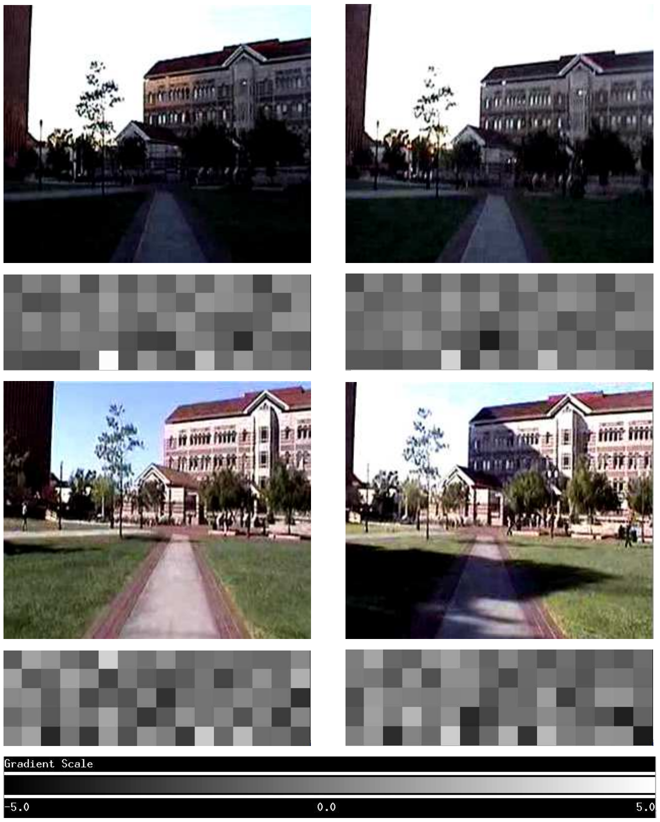


Fig. 15. Lighting Conditions use for Testing at Frederick D. Fagg park (FDF). Clockwise from top left: early evening, evening, noon, and middle of afternoon

TABLE VI
FREDERICK D. FAGG PARK CONFUSION MATRIX

Segment number guessed by algorithm

True segment number

Segment	1	2	3	4	5	6	7	8	9
1	2641	11	165	0	0	364	0	170	0
2	55	2796	106	0	7	97	0	70	72
3	0	71	3111	0	0	153	0	0	11
4	0	136	1	3006	0	0	0	92	0
5	9	23	0	0	2349	63	13	660	10
6	22	30	254	5	6	6288	0	3	21
7	0	154	0	0	24	62	3256	168	47
8	45	30	40	35	6	3	40	4613	196
9	3	5	68	56	4	38	2	224	2701

increased by 1,800 (18 new output nodes by 100 second hidden-layer nodes), from 36,900 to 38,700 connections (4.88%). We use the same procedure as well as training and testing data (175,406 and 75,073 frames, respectively). The training process takes much longer than for the other experiments. It is about 260 epochs with the last 200 epochs or so converging very slowly from 3% down to 1%. When training, we print the confusion matrix periodically to analyze the process of convergence, and we find that the network converges from inter-site classification before going further and eliminate the intra-site errors. We organize the results into segment-level (Table VII) and site-level (Table VIII) statistics.

For segment-level classification, the total error rate is 13.55%. We expected the results to be somewhat worse than all previous three experiments when each site is classified individually. However, such is not the case when comparing it with the AnF experiment (15.79%; experiment 2) while being marginally worse than the other two (12.04% for ACB in experiment 2 and 11.38% for FDF in experiment 3). Notice also that the relative error among the individual sites changes as well. The results for AnF segments in the combined setup improve by 2.47% to 13.32% error while rate for segments

TABLE VIII
COMBINED SITES EXPERIMENTAL RESULTS

Site	ACB	AnF	FDF	False-/Total	Pct. err
ACB	12882	563	520	1083/13965	7.76%
AnF	350	25668	379	729/26397	2.76%
FDF	163	1433	33115	1596/34711	4.60%
False+	513	1996	899	3408	
Total	13395	27664	34014	75073	
Pct. err	3.83%	7.22%	2.64%	4.54%	

in ACB and FDF degrades by 4.24% and 1.25%, respectively. From the site-level confusion matrix (table VIII), we see that the system can reliably pin a given test image to the correct site with only 4.54% error (95.46% classification). This is encouraging because the classifier can then provide various levels of outputs. For instance, when the system is unsure about the actual segment location, it can at least rely on being at the right site.

One of the concerns in combining segments from different sites is that the number of samples for each of them becomes unbalanced as there are some segments that take less time to walk through (as short as 10 seconds) while others can take up to a minute and a half. That is, the lower number of samples for ACB may yield a network convergence that gives heavier weights on correctly classifying the longer segments from AnF and FDF. From the site-level statistics (table VIII), we can see that the trend does hold, although not to an alarming extent.

IV. DISCUSSION

We have shown that the gist features succeed in classifying a large set of images without the help of temporal filtering (one-shot recognition on each image considered individually), which would be expected to further improve the results by reducing noise significantly [17]. In terms of robustness, the features are able to handle translational, angular, scale and illumination changes. Because they are computed from large image sub-regions, it takes a sizable translational shift to affect the values. As for angular stability, the natural perturbation of a camera carried while walking seems to aid the demonstrated invariance. For larger angular discrepancy, like for example in in the case of off-road environments, an engineering approach like adding sensors such as a gyroscope to correct the angle of view may be advisable. The gist features are also invariant to scale because the majority of the scenes (background) are stationary and the system is trained at all viewing distances. The gist features achieve a solid illumination invariance when trained with different lighting conditions. Lastly, the combined-sites experiment shows that the number of differentiable scenes can be quite high. Twenty seven segments can make up a detailed map of a large area.

A profound effect of using gist is the utilization of background information moreso than foreground. However, one drawback of the current gist implementation is that it cannot carry out partial background matching for scenes in which large parts are occluded by dynamic foreground objects. As mentioned earlier the videos are filmed during off-peak hours when few people (or vehicles) are on the road. Nevertheless, they can still create problems when moving too close to the camera. In our system, these images can be taken out using the motion cues from the motion channel of the saliency algorithm as a preprocessing filter, detecting significant occlusion by thresholding the sum of the motion channel feature maps [37]. Furthermore, a wide-angle lens (with software distortion correction) can help to see more of the background scenes and, in comparison, decrease the size of the moving foreground objects.

The gist features, despite their current simplistic implementation, are able to achieve a promising localization performance. The technique highlights the rapid nature of gist while still accurate in performing its tasks. This, in large part, is because the basic computational mechanisms of extracting the gist features are simple averaging of visual cues from different domains. Theoretically, scalability is a concern because when we average over large spaces, background details that may be critical in distinguishing certain locations can be lost. However, although more sophisticated gist computation could be incorporated, we avoid complications that occur when trying to fit more complex models to unconstrained and noisy data. For example, a graph would be a more expressive layout representation than the current grid-based decomposition (refer to figure 3). It can represent a scene as segmented region feature vectors, or even objects, for each node, and coarse spatial relationships for the edges. The node information can provide explicit shape recognition, which, in essence, is what is lacking in our current implementation. Howver, as mentioned

TABLE VII
COMBINED SEGMENT EXPERIMENTAL RESULTS

	ACB				AnF				FDF			
Segment	False+	% err.	False-	% err.	False+	% err.	False-	% err.	False+	% err.	False-	% err.
1	292/1657	16.20%	231/1596	14.47%	565/3120	18.11%	582/3137	18.55%	231/2306	10.02%	1276/3351	38.08%
2	277/1710	16.20%	342/1775	19.27%	636/2159	29.46%	323/1846	17.50%	455/3175	14.33%	483/3203	15.08%
3	275/2031	13.54%	130/1886	6.89%	555/4198	13.22%	130/3773	3.45%	893/3881	23.01%	358/3346	10.70%
4	61/1211	5.04%	84/1234	6.81%	233/2401	9.70%	237/2405	24.05%	56/3102	1.81%	189/3235	5.84%
5	129/1208	10.68%	205/1284	15.97%	583/3251	17.93%	270/2938	9.19%	107/3115	3.43%	119/3127	3.81%
6	162/2040	7.94%	209/2087	10.01%	926/4462	20.75%	688/4224	16.29%	784/6426	12.20%	987/6629	14.89%
7	308/1438	21.42%	551/1681	32.78%	298/1680	17.74%	570/1952	29.20%	309/3704	8.34%	316/3711	8.52%
8	83/961	8.64%	221/1099	20.11%	730/3278	22.72%	307/2855	10.75%	300/4833	6.21%	475/5008	9.48%
9	116/1139	10.18%	300/1323	22.68%	257/3115	8.25%	409/3267	12.52%	551/3472	15.87%	180/3101	5.80%
Total	1703/13395	12.714%	2273/13965	16.276%	4783/27664	17.290%	3516/26397	13.320%	3686/34014	10.837%	4383/34711	12.627%
Total	10172/75073 = 13.55%											

in introduction such approach can break down when a segmentation error occur. In our second experiment (AnF), for example, overlapping trees or overlapping buildings can be jumbled together.

Another way to increase the theoretical strength of the gist features is to go to a finer grid to incorporate more spatial information. For the current extraction process, to go to next level in the pyramid (an eight-by-eight grid) is to increase the number of features from 16 to 64 in each sub-channel. However more spatial resolution also means more data (quadrupled the amount) to process and it is not obvious where the point of diminishing return is. We have to strike a balance in resolution and generalization, pushing the complexity of expressiveness of the features while keeping robustness and compactness in mind. That said, our goal is to emulate human-level gist understanding that can be applied to a larger set of problems. As such our further direction would be to stay faithful to the available scientific data on human vision.

We have not discussed in length the need for localization within a segment. The gist features would have problems differentiating scenes when most of the background overlaps as is the case for scenes within segment. Gist, by definition, is not a mechanism to produce a detailed picture and an accurate localization, just the coarse context. This is where saliency and localized object recognition may complement gist. The raw gist features are shared with the saliency model so that we can attack the problem from multiple sides efficiently. One of the issues we encounter with this arrangement will be the need to synchronize the representation so that the two can coincide naturally to provide a complete scene description. With gist we can locate our general whereabout to the segment level. With saliency we can now pin-point our exact locations by finding distinctive cues situated within the segment and approximate our distance from them. The gist model can even prime regions within a scene [46] by providing context to prune out noisy possible salient locations.

V. Conclusion

We have shown that the gist model can be useful in outdoors localization for a walking human, with obvious application to autonomous mobile robotics. The model is able to provide high-level context information (a segment within a site) from various outdoor environments despite using coarse features. It is able to contrast scenes in a global manner and automatically takes obvious idiosyncrasies into account. This capability reduces the need for detailed calibration in which a robot has to rely on the ad-hoc knowledge of the designer for reliable landmarks. Furthermore, we are working on extending the current system to recognize places/segments without having to train it explicitly. This requires the ability to cluster gist feature vectors from a same location, which also help alert the robot when it is moving from one location to another.

Because the raw features are shared with the saliency model, the system can efficiently increase localization resolution. It can use salient cues to create distinct signatures of individual scenes, finer points of reference within a segment that may not be differentiable by gist alone. The salient cues can also help guide localization for the transitions between segments which we did not try to classify.

In the future we would like to present a physical implementation of a model that uses bottom-up salient cues as well as context to produce a useful topographical map for navigation in unconstrained, outdoor environment.

ACKNOWLEDGMENT

This research was supported by the National Geospatial-Intelligence Agency, the National Science Foundation, and the Human Frontier Science Program. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

REFERENCES

[1] D. Fox, W. Burgard, F. Dellaert, and S. Thrun, "Monte carlo localization: Efficient position estimation for mobile robots," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI'99).*, July 1999.

[2] J. J. Leonard and H. F. Durrant-Whyte, "Mobile robot localization by tracking geometric beacons," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 3, pp. 376–382, June 1991.

[3] S. Thrun, D. Fox, and W. Burgard, "A probabilistic approach to concurrent mapping and localization for mobile robots," *Machine Learning*, vol. 31, pp. 29–53, 1998.

[4] K. Lingemann, H. Surmann, A. Nuchter, and J. Hertzberg, "Indoor and outdoor localization for fast mobile robots," in *IROS*, 2004.

[5] Y. Abe, 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*, vol. 20, pp. 1299–1304, May 1999.

[6] S. Thrun, "Finding landmarks for mobile robot navigation," in *ICRA*, 1998, pp. 958–963.

[7] S. Maeyama, A. Ohya, and S. Yuta, "Long distance outdoor navigation of an autonomous mobile robot by playback of perceived route map," in *ISER*, 1997, pp. 185–194.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] S. Se, D. G. Lowe, and J. J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Transactions on Robotics*, vol. 21, no. 3, pp. 364–375, 2005.

[10] L. Goncalves, E. D. Bernardo, D. Benson, M. Svedman, J. Ostrowski, N. Karlssona, and P. Pirjanian, "A visual front-end for simultaneous localization and mapping," in *ICRA*, April 18 - 22 2005, pp. 44–49.

[11] H. Katsura, J. Miura, M. Hild, and Y. Shirai, "A view-based outdoor navigation using object recognition robust to changes of weather and seasons," in *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robot and Systems (IROS 2003)*, Las Vegas, Nevada, US, October 27 - 31 2003, pp. 2974–2979.

[12] Y. Matsumoto, M. Inaba, and H. Inoue, "View-based approach to robot navigation," in *IEEE-IROS*, 2000, pp. 1702–1708.

[13] R. Murrieta-Cid, C. Parra, and 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, 2002.

[14] L. Renniger and J. Malik, "When is scene identification just texture recognition?" *Vision Research*, vol. 44, pp. 2301–2311, 2004.

[15] I. Ulrich and I. Nourbakhsh, "Appearance-based place rcognition for topological localization," in *IEEE-ICRA*, April 2000, pp. 1023 – 1029.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[17] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France, October 2003, pp. 1023 – 1029.

[18] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97 – 137, 1980.

[19] J. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin and Review*, vol. 1, no. 2, pp. 202 – 238, 1994.

[20] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[21] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[22] M. C. Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, 1975.

[23] I. Biederman, "Do background depth gradients facilitate object identification?" *Perception*, vol. 10, pp. 573 – 578, 1982.

[24] B. Tversky and K. Hemenway, "Categories of the environmental scenes," *Cognitive Psychology*, vol. 15, pp. 121 – 149, 1983.

[25] A. Oliva and P. Schyns, "Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli," *Cognitive Psychology*, vol. 34, pp. 72 – 107, 1997.

[26] T. Sanocki and W. Epstein, "Priming spatial layout of scenes," *Psychol. Sci.*, vol. 8, pp. 374 – 378, 1997.

[27] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, pp. 17 – 42, 2000.

[28] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520 – 522, 1995.

[29] M. M. MJ, S.J.Thorpe, and M. Fabre-Thorpe, "Rapid categorization of achromatic natural scenes: how robust at very low contrasts?" *Eur J Neurosci.*, vol. 21, no. 7, pp. 2007 – 2018, April 2005.

[30] F. Li, R. VanRullen, C. Koch, and P. Perona, "Rapid natural scene categorization in the near absence of attention," in *Proc. Natl. Acad. Sci.*, 2002, pp. 8378 – 8383.

[31] R. Epstein, D. Stanley, A. Harris, and N. Kanwisher, "The parahippocampal place area: Perception, encoding, or memory retrieval?" *Neuron*, vol. 23, pp. 115 – 125, 2000.

[32] A. Oliva and P. Schyns, "Colored diagnostic blobs mediate scene recognition," *Cognitive Psychology*, vol. 41, pp. 176 – 210, 2000.

[33] A. Torralba, "Modeling global scene factors in attention," *Journal of Optical Society of America*, vol. 20, no. 7, pp. 1407 – 1418, 2003.

[34] C. Ackerman and L. Itti, "Robot steering with spectral image information," *IEEE Transactions on Robotics*, vol. 21, no. 2, pp. 247–251, Apr 2005.

[35] L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of Visual Behavior*, D.J.Ingle, M.A.Goodale, and R.J.W.Mansfield, Eds. Cambridge, MA: MIT Press, 1982, pp. 549–586.

[36] L. Itti and C. Koch, "A saliency-based search mechanism for

overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10-12, pp. 1489–1506, May 2000.

[37] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct 2004.

[38] ——, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, California Institute of Technology, Pasadena, California, Jan 2000.

[39] R. S. Turner, *In the eye's mind : vision and the Helmholtz-Hering controversy*. Princeton University Press, 1994.

[40] C. Chubb and G. Sperling, "Drift-balanced random stimuli: a general basis for studying non-fourier motion perception," *JOSA*, vol. 5, no. 11, pp. 1986 – 2007, 1988.

[41] K. Barnard, V. Cardei, and B. Funt, "A comparison of computational color constancy algorithms; part one: Methodology and experiments with synthesized data," *IEEE Transactions in Image Processing*, vol. 11, no. 9, pp. 972 – 984, 2002.

[42] K. Barnard, L. Martin, , A. Coath, and B. Funt, "A comparison of color constancy algorithms. part two. experiments with image data," *IEEE Transactions in Image Processing*, vol. 11, no. 9, pp. 985 – 996, 2002.

[43] G. Finlayson, B. Schiele, and J. Crowley, "Comprehensive colour image normalization," in *5th European Conference on Computer Vision*, May 1998, pp. 475 – 490.

[44] A. Hyvrinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[45] C. Siagian and L. Itti, "Gist: A mobile robotics application of context-based vision in outdoor environment," in *Proc. IEEE-CVPR Workshop on Attention and Performance in Computer Vision (WAPCV'05), San Diego, California*, Jun 2005, pp. 1–7.

[46] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *IEEE Proc. Of Int. Conf in Comp. Vision*, 2001, pp. 763 – 770.

Christian Siagian is currently working towards a Ph.D. degree in the field of Computer Science. His research interests include robotics and computer vision, such as vision-based mobile robot localization and scene classification, particularly the ones that are biologically-inspired.

Laurent Itti received his M.S. degree in Image Processing from the Ecole Nationale Supérieure des Télécommunications in Paris in 1994, and his Ph.D. in Computation and Neural Systems from Caltech in 2000. He is now an associate professor of Computer Science, Psychology, and Neuroscience at the University of Southern California. Dr. Itti's research interests are in biologically-inspired computational vision, in particular in the domains of visual attention, gist, saliency, and surprise, with technological applications to video compression, target detection, and robotics.