massachusetts institute of technology — artificial intelligence laboratory

Recognizing Indoor Scenes

Antonio Torralba and Pawan Sinha

AI Memo 2001-015 July 2001
CBCL Memo 202

Abstract

We propose a scheme for indoor place identification based on the recognition of global scene views. Scene views are encoded using a holistic representation that provides low-resolution spatial and spectral information. The holistic nature of the representation dispenses with the need to rely on specific objects or local landmarks and also renders it robust against variations in object configurations. We demonstrate the scheme on the problem of recognizing scenes in video sequences captured while walking through an office environment. We develop a method for distinguishing between 'diagnostic' and 'generic' views and also evaluate changes in system performances as a function of the amount of training data available and the complexity of the representation.

This research was supported in part by funds from the Alfred P. Sloan Fellowship for Neuroscience to P.S., the Defense Advanced Research Projects Agency and HONDA R&D. The authors wish to thank Dr. Aude Oliva for fruitful discussions.

I. Introduction

Much attention in high-level vision has been devoted to the problem of individual object recognition. An equally important but less researched problem is that of recognizing entire scenes. Scene recognition underlies many other abilities, most notably navigation through complex environments.

Scene-based navigation strategies are prevalent in the animal kingdom. Animals as simple as bees and ants can perform impressive feats of navigation using image matching. Desert ants of the genus Cataglyphis, for example, can reliably return to their nest following foraging trips that can exceed several hundred meters in length. [Wehner, 1987]. It is unlikely that Cataglyphis use chemical cues for path guidance since no marker would persist long enough in the extreme heat of the desert. Careful experiments suggest that the ants rely primarily on the visual information regarding the environment [Wehner and Raber, 1979]. There is evidence that bees too associate direction vectors with particular scenes [Cartwright and Collett, 1987]. When captured near their hives and transported in enclosed containers to different feeding spots, bees upon being released are able to fly straight in the direction of the hive. Since they cannot rely on dead-reckoning in these experiments, the likely explanation is that the bees learn to associate specific direction vectors with particular scenes. Such associations are useful when insects have to learn complex foraging routes and are required to execute a sequence of vectors in the correct order [Janzen, 1971].

Most of systems developed for localization of robotic systems based on visual information focus on the analysis of 3D scene information and/or the location of visual landmarks like edges or interest points [see Borenstein et al, 1996 for a review]. A different approach for localization is used by research in wearable computing [e.g. Clarkson et al, 2000] in which the system uses information about the statistics of simple sensors (acoustic and visual) for identifying coarse locations and events.

Besides navigation, many other perceptual abilities such as object localization also rely on scene recognition. This, in general, is a complex task. One way to reduce the complexity of the problem is by relying on prominent landmarks or distinctive markings in the environment. However, such localized cues may not always be readily available in all circumstances. A general-purpose scene recognition scheme has to be able to function without critically relying on distinctive objects. In this paper, we develop a system to accomplish this task in arbitrary indoor environments. Our scheme represents scene structure holistically [Oliva and Torralba, 2001] and, therefore, does not require the presence of specific landmarks.

II. Low dimensional scene representation

Much of the prior work on scene recognition uses the identities of specific objects present in a scene for scene classification. However, this strategy requires a prior step of object recognition. Furthermore, the human visual system is able to analyze scenes even under degraded conditions that obscure the identities of individual objects [Schyns and Oliva, 1994]. We therefore opt to develop a holistic representation of scene structure that does not need a prior assessment of individual objects. The overall scene properties that are believed to be relevant for discriminating between different scenes are (e.g. Gorkani and Picard, 1994; Carson et al, 1997; Lipson et al, 1997; Oliva and Torralba, 2001; Szummer and Picard, 1998; Torralba and Oliva, 2000; Vailaya et al, 1998; De Bonet and Viola, 1997):

- The statistics of structural elements: Different structural elements (e.g., buildings, road, tables, walls, with particular orientation patterns, smoothness/roughness) compose each context (e.g., rooms, streets, shopping center).
- The spatial organization: The structural elements have particular spatial arrangements. Each context imposes certain organization laws (e.g. for streets: road in the bottom, buildings in the sides, an aperture in the center).
- Color distribution

As described below, we use a low dimensional holistic representation that encodes the structural scene properties. Color is not taken into account in this study, although the framework can be naturally extended to include this attribute. The image features most commonly used for describing local structures are the energy outputs of oriented band-pass filters, as they have been shown to be relevant for the task of object detection [e.g. Itti et al, 1998; Rao et al, 1996; Schiele and Crowley, 1997] and scene recognition [e.g. Gorkani and Picard, 1994; Oliva and Torralba, 2001]. Therefore, the local image representation at the spatial location $(\vec{x})$ is given by the vector $\vec{v}_L(\vec{x}) = \{v(\vec{x}, k)\}_{k=1,N}$ with:

$$v(\vec{x}, k) = \left| \sum_{\vec{x}'} i(\vec{x}') g_k(\vec{x} - \vec{x}') \right| \tag{1}$$

$i(\vec{x})$ is the input image and $g_k(\vec{x})$ are oriented band-pass filters defined by $g_k(\vec{x}) = e^{\|\vec{x}\|^2/\sigma_k^2} e^{2\pi j < \vec{f_k}, \vec{x} >}$. In such a representation, $v(\vec{x}, k)$ is the output magnitude at the location $\vec{x}$ of a complex Gabor filter tuned to the spatial frequency $\vec{f_k}$. The variable k indexes filters tuned to different spatial frequencies and orientations. The absolute value provides some invariance with respect to the input phase information.

In order to reduce the dimensionality of this representation, we decompose the image features $v(\vec{x}, k)$ into the basis functions provided by PCA

$$a_n = \sum_{\vec{x}} \sum_{k} v(\vec{x}, k)\, \psi_n(\vec{x}, k) \tag{2}$$

with:

$$v(\vec{x}, k) \simeq \sum_{n=1}^{D} a_n \psi_n(\vec{x}, k) \tag{3}$$

We propose to use the decomposition coefficients $\vec{v}_C = \{a_n\}_{n=1,D}$ as context features. The functions $\psi_n(\vec{x}, k)$ are the eigenfunctions of the covariance operator given by $v(\vec{x}, k)$. Therefore, the functions $\psi_n(\vec{x}, k)$ incorporate both spatial and spectral information. D is the dimensionality of the representation. By using only a reduced set of components ($D = 60$ for the rest of the paper), the coefficients $\{a_n\}_{n=1,D}$ encode the main spectral characteristics of the scene with a coarse description of their spatial arrangement. In essence, $\{a_n\}_{n=1,D}$ is a holistic representation as all the regions of the image contribute to all the coefficients, and objects are not encoded individually [see Oliva and Torralba, 2001].

As shown in figure 1 the first principal components encode only low resolution spatial and spectral information. The low-resolution representation, combined with the absolute value in 1, provides some robustness with respect to objects arrangements. This is an important factor for scene representation as particular scenes are defined by the coarse organization of the major elements (bookshelves, tables, doors, windows, etc.) without being affected by the redistribution of minor elements as office supplies, chairs, books, etc.

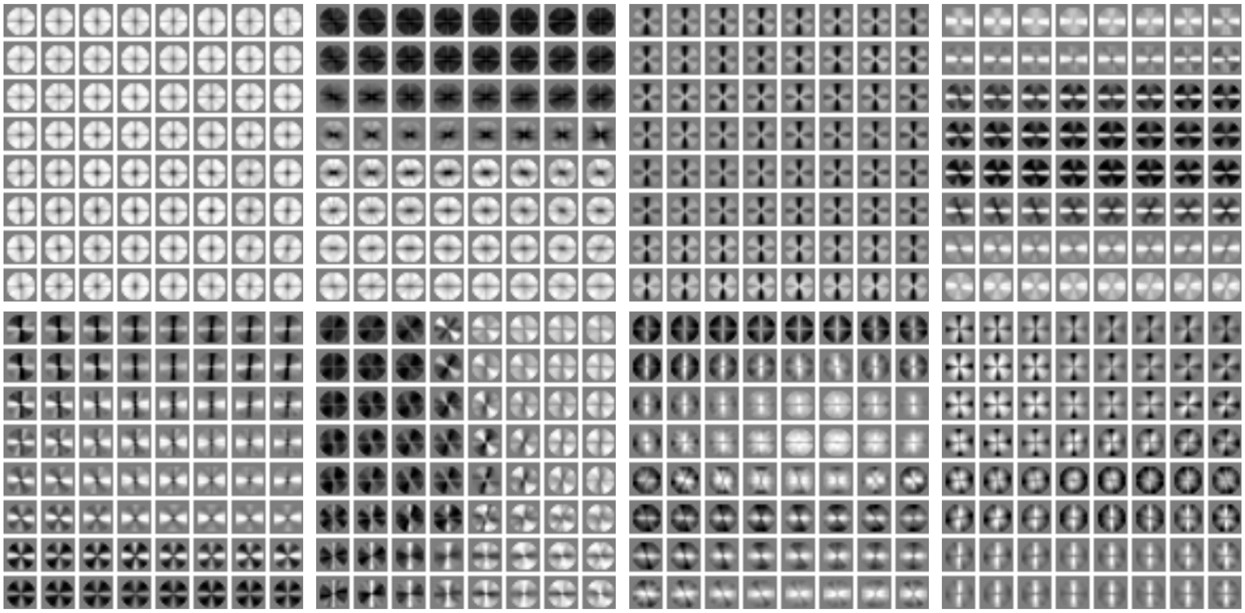


Fig. 1. Examples of functions $\psi_n(\vec{x}, k)$. For simplicity of the visualization we show the principal components of the spectrogram [Oliva and Torralba, 2001]. The figure shows how the spectral components are weighted at each spatial location to obtain a_n.

This representation has been shown to be relevant for outdoor natural and urban scene categorization [Oliva and Torralba, 2001] and for modeling contextual influences on object detection and recognition [Torralba and Sinha, 2001]. We show next that it is also effective for recognizing indoor scenes.

III. Visual scene landmarks

When trying to identify a place based on a single view, not all-possible points of view will provide enough information for making a reliable decision. For instance, in the case of a robot exploring the environment, many of the views may be close-up views of simple surfaces or views of generic objects (fig. 5.b). Such frames are likely to be ambiguous as similar views may be found in many different places. The views that provide useful information for place identification will depend on the number of different places to discriminate and the variability among places (fig. 5.a).

In the system that we propose, place recognition will be based on *visual scene landmarks.* This is distinct from visual landmarks based on particular objects or signs. Our system will recognize global views of the environment without the use of localized information

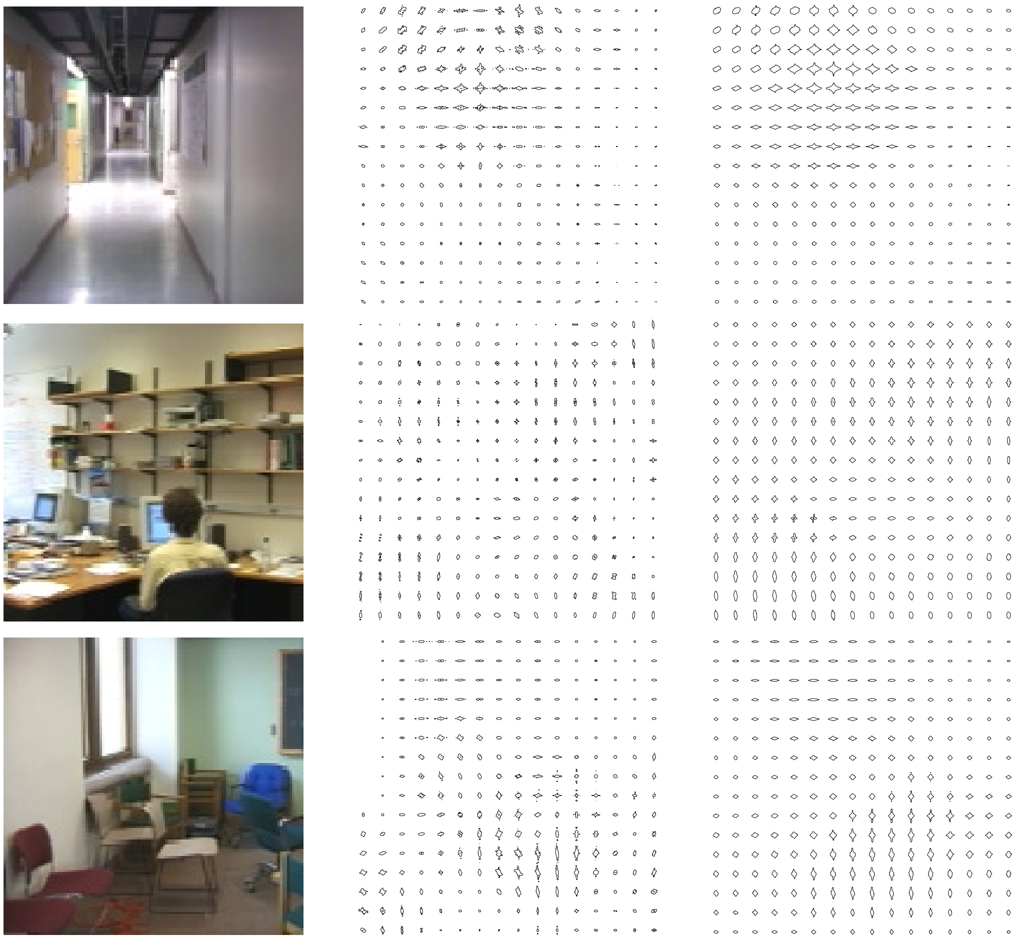


Fig. 2. Spatial layout of spectral components. The middle image shows the section of the magnitude of the local Fourier transforms (spectrogram) for each image. The right hand side image shows the approximation of spectrogram obtained from the first 20 PCs. The obtained layout captures the dominant orientations and scales with coarse image regions.

as objects. Each view is encoded using the holistic structural features described in the preceding section. Therefore, the system does not require building a 3D model of the scene.

Figure 3 shows a few scenes of a film that simulates a visit to the Department of Brain and Cognitive Science at MIT. The total film contains about 3000 frames (5 frames per second) and 15 different places and is used for training the system. The film was made without taking any particular care to choose good viewpoints. Some of the views are close-ups of doors, walls or objects and do not provide information about the identity of the place, other views are global views of the place. Some frames are corrupted by motion

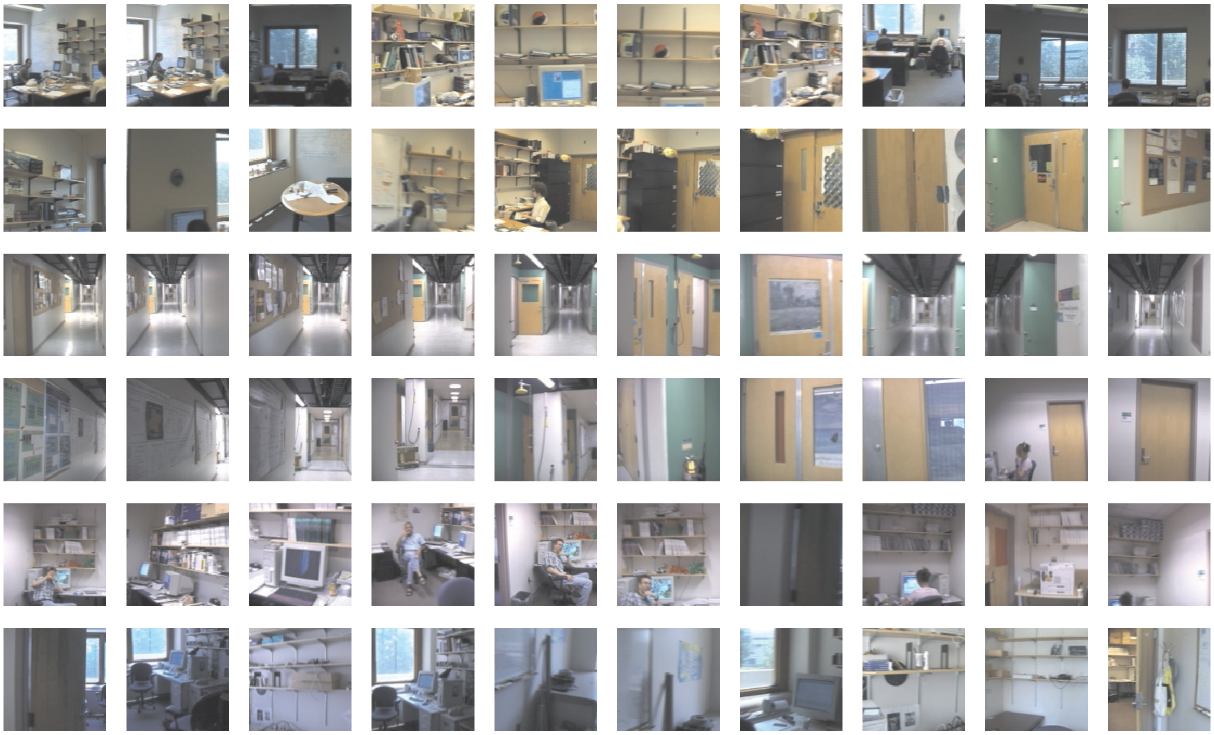


Fig. 3. Example of some frames of a sequence taken while visiting the lab.

blur or have noise due to poor illumination. Those frames were not removed from the sequence.

The film used for training entailed visiting the different places and trying to avoid loops so that the time in the sequence is roughly correlated with distance from the departure point. By avoiding repeat visits in the training video, we implicitly have a way for distinguishing between non-generic and generic views. Non-generic views (*landmarks*) would not be expected to have high correlations with frames beyond one contiguous segment of the video. Generic views, on the other hand, will yield high correlations across multiple segments because they are not specific to any particular place.

One way of accounting for the similarities of a single frame across the temporal sequence is by means of the conditional probability density function $p(t \,|\, \vec{v}_C)$. Given the structural features $\vec{v}_C$ of a frame, the PDF provides the distribution of frames that have similar features. The PDF $p(t \,|\, \vec{v}_C)$ represents the intuitive notion of 'when did I see a scene similar to this one'. $p(t \,|\, \vec{v}_C)$ provides the most likely temporal window in the training sequence that can produce the structural features $\vec{v}_C$.

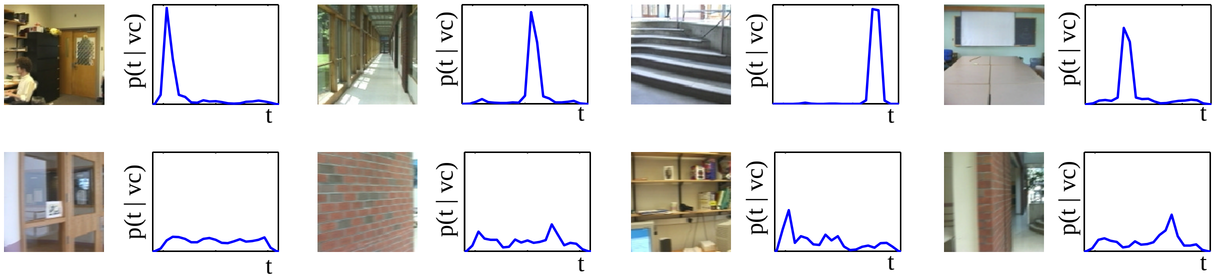


Fig. 4. The conditional PDF $p(t \,|\, \vec{v}_C)$ has very different forms for distinctive versus generic views and can therefore be used to select discriminant frames.

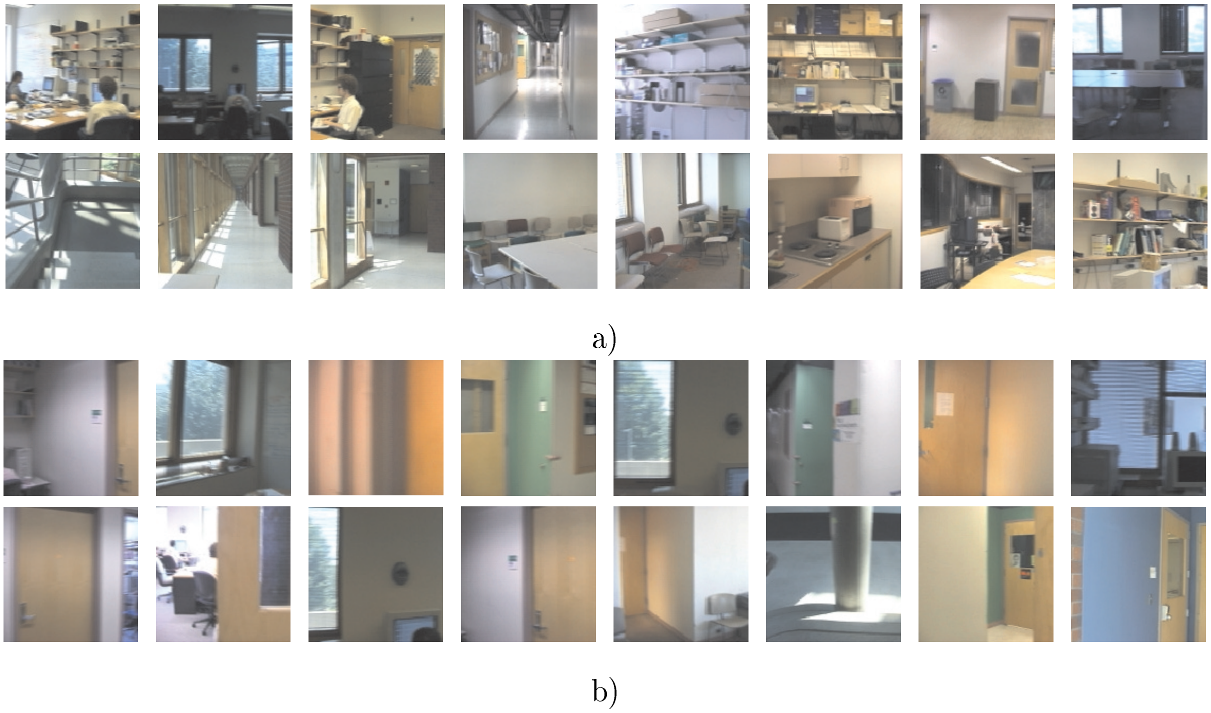
The PDF $p(t \,|\, \vec{v}_C)$ can be modeled by a Parzen window approach. For each frame of the training film we compute the structural features presented before. Then, the training data set consists of the pairs t_i (the time index of the frame i in the training film) and the structural features corresponding to the frame $\vec{v}_i$, with $i = 1, ..., N_f$, being N_f the number of frames of the film. Then, given a new frame not included in the training, the PDF $p(t \,|\, \vec{v}_C)$ is:

$$p(t \,|\, \vec{v}_C) = \frac{\sum_{i=1}^{N_f} K_{\sigma_1}(t - t_i) K_{\sigma_2}(\vec{v}_C - \vec{v}_i)}{\sum_{i=1}^{N_f} K_{\sigma_2}(\vec{v}_C - \vec{v}_i)} \tag{4}$$

The kernel $K_\sigma(\vec{x}) = k\, exp(-\|\vec{x}\|^2/\sigma^2)$ is a radial gaussian kernel with width σ^2. k is a normalization constant so that the kernel averages to one. In order to have a better estimate of $p(t \,|\, \vec{v}_C)$ we average the PDF obtained during N consecutive frames. N is selected so that the views considered belong to the same place and have relatively small variations with respect to the point of view (here $N = 10$). The kernel $K_{\sigma_2}(\vec{v}_C - \vec{v}_i)$ accounts for the similarity between the target frame and the training sequence (the target frame is not included in the training sequence used to evaluate eq. 4). If the target frame has good matches in the sequence, then the PDF $p(t \,|\, \vec{v}_C)$ will have one or few maximums that will indicate at which times in the training sequence there are frames similar to the target frame. If there are no good matches, then the PDF $p(t \,|\, \vec{v}_C)$ will be a uniform distribution (fig. 4).

The conditional entropy H provides a simple way for accounting for the dispersion:

$$H = -\int p(t \,|\, \vec{v}_C) \log p(t \,|\, \vec{v}_C) dt \tag{5}$$



a)

b)

Fig. 5. a) Visual landmarks and b) generic views. The two categories have been derived using the conditional entropy measure.

Figure 5.a shows a collection of views with small H and figure 5.b shows views with the largest H. As expected, images with low H correspond to large views of the environment providing enough information for identifying the place. Views with large H correspond to close-up views, doors, windows and walls which provide ambiguous information for place identification.

Until this point, no information about the identity of each place has been given to the system. Place identification requires that the training film also provide the identity of each place at each frame (fig. 6).

IV. Place identification

A. Training

The system was trained to differentiate among 15 different places (fig. 7). The training is performed using one video sequence for which we provide the system with the identity of each place at each instant by labeling the different time windows in which the camera is recording one place (fig. 6). The sequence contains a large variety of points of view for

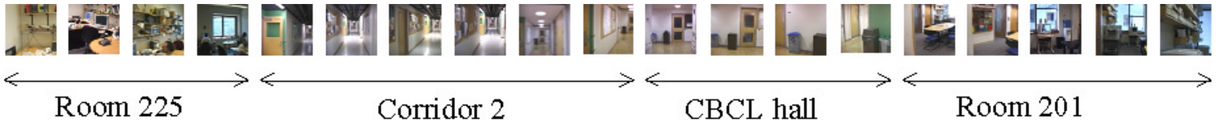


Fig. 6. The annotated training set for the system is created by labeling segments of the video sequence.

each place.

For the recognition, we evaluate the probabilities $p(C_i|\vec{v}_C)$ where $\vec{v}_C$ are the structural features of an image that we want to identify and C_i are the labels of the 15 categories defined. If we assume that the places C_i represent all possible places that the robot can be in, then the PDFs $p(C_i|\vec{v}_C)$ are modeled by:

$$p(C_i|\vec{v}_C) = \frac{p(\vec{v}_C|C_i)p(C_i)}{p(\vec{v}_C)} = \frac{p(\vec{v}_C|C_i)p(C_i)}{\sum_j p(\vec{v}_C|C_j)p(C_j)} \tag{6}$$

with:

$$p(\vec{v}_C|C_i) = \sum_{j \in C_i} K_\sigma(\vec{v}_C - \vec{v}_j) \tag{7}$$

where the kernel $K_\sigma(\vec{x})$ is a radial gaussian kernel with width σ. The parameters of the PDF $p(C_i|\vec{v}_C)$ are the structural features $\vec{v}_j$ of the images of the training sequence that correspond to each place. We set $p(C_i) = 1/15$.

B. *Results*

For testing the ability of the holistic representation to discriminate among the 15 places in which the system was trained, we used new sequences recorded on different days and times.

For each frame of the new sequences, we assign a label for the category that provides the maximum $p(C_i|\vec{v}_C)$ but only if this probability is above a predefined threshold. If the maximum probability value is not high enough, then the frame is not labeled. The probability $p(C_i|\vec{v}_C)$ provides a measure of the confidence that the system has for assigning the label C_i to one frame. The frames that correspond to scene landmarks will produce high confidence classifications. On the other hand, generic views will match several places and the probabilities $p(C_i|\vec{v}_C)$ will remain under the threshold. Figure 8 shows some examples of the $p(C_i|\vec{v}_C)$ obtained with frames from a new sequence. The bars on the right of each

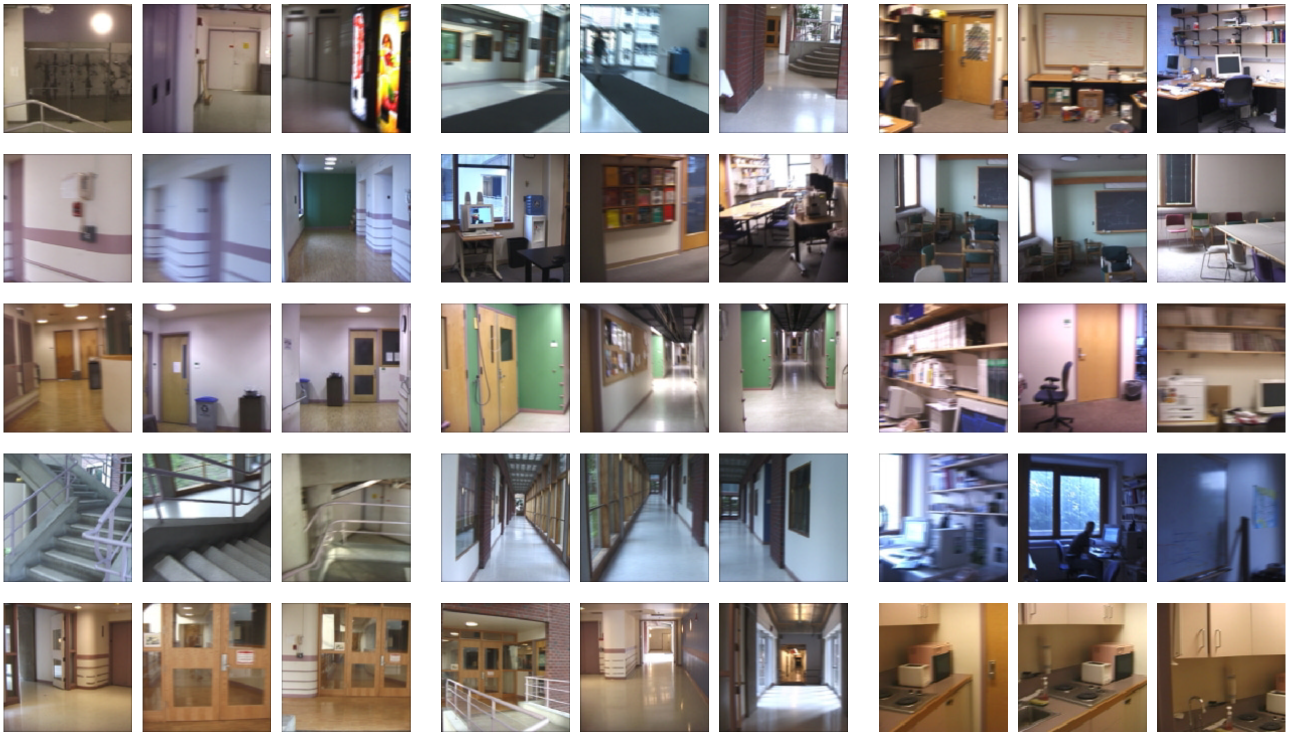


Fig. 7. Examples of the 15 places (three pictures per place) that the system has to recognize.

image indicate the likelihood of the image belonging to each of the 15 possible classes. Whenever any likelihood value exceeds a threshold (indicated here by the horizontal line) the system outputs the corresponding label. The three images at the top correspond to high confidence recognized frames (visual scene landmarks) and the three images in the bottom are unlabeled frames and correspond to generic views.

By using a high threshold (near 1), the system will require a high confidence to put a label and therefore most of the time the system will not take any decision. By using a low threshold, the system will take decisions almost in any frame increasing the number of errors.

Therefore, there is a trade off between the percentage of time that we need the system to take decisions and the accuracy of performance. This trade off is also a function of a few other parameters of the system, in particular:

- N: The number of frames used for the training of the PDF $p(C_i|\vec{v}_C)$ for each place.
- D: The dimensionality of the scene representation. Increasing the dimensionality increases the distinctiveness of each scene.

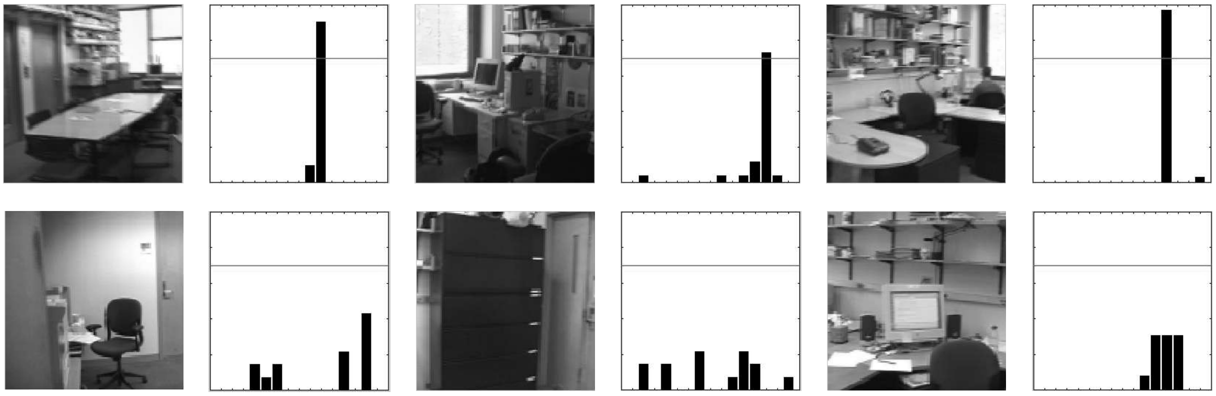


Fig. 8. Some results. The bars on the right of each image indicate the likelihood of the image belonging to each of the 15 possible classes. Whenever any likelihood value exceeds a threshold (indicated here by the horizontal line) the system outputs the corresponding label. The three images at the top correspond to high confidence recognized frames (visual scene landmarks) and the three images in the bottom are unlabeled frames and correspond to generic views.

- T: The extend of temporal integration: To make the system more robust we integrate the obtained probabilities $p(C_i|\vec{v}_C)$ over time by averaging over T frames (for the results presented here, $T = 10$ frames).

Figure 9 summarizes the results obtained demonstrating the influence of changes in N and D. By decreasing the number of frames used for training we decrease the quality of the estimation of the PDF $p(C_i|\vec{v}_C)$ and, therefore, performance degrades (fig. 9.a). When using less than 25 training images per category, the system fails to reach high confidence ratings for nearly all frames.

Fig. 9.b shows the results when reducing the dimensionality of the representation (the training set size is 100 images per category for all the graphs). By reducing the number of features we decrease the amount of time that the system can be confident about the identity of the place. But, even for very low dimensional representations, the system can still provide a few labels with high confidence and yielding a 95% recognition rate within this set of labeled frames.

Figure 10 shows the results of place identification in a sequence in which the confidence level was set experimentally in order to provide labels at least for 40% of the frames.

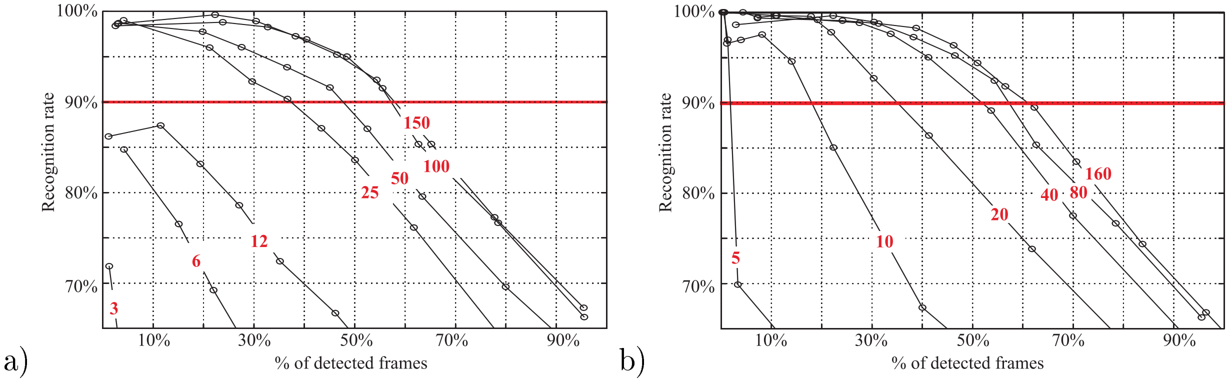


Fig. 9. Each curve shows system performance as a function of the proportion of frames that the system is required to label. This is adjusted by varying the degree of confidence required to take a decision. By lowering the confidence requirements we increase the number of frames with label but in doing so we decrease performance. The two graphs show the performance as a function of the number of training images for each category (a) and as a function of the number of structural features used for representing each frame (b). The best trade-off between model complexity and performance is obtained with 80 features per image and 100 images for the training set. However, the computational complexity can be reduced if the application allows providing labels in a smaller percentage of frames.

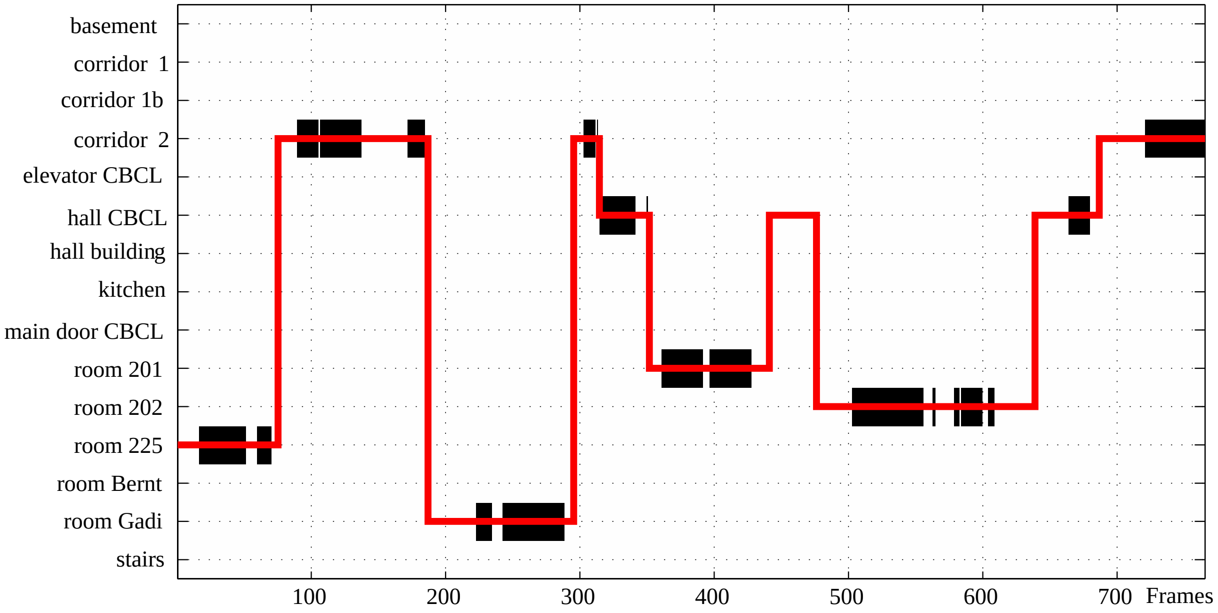


Fig. 10. Some results. 100% correct detected in this sequence. The continuous line indicates the ground truth of the places visited in the sequence. The thick lines indicates the labeled frames.

V. Conclusion

The problems of individual object recognition on the one hand and scene recognition on the other have been treated as qualitatively different undertakings. Past studies have typically treated scene classification as a problem of inference that obtains its basic tokens from a prior step of object recognition. The viewpoint we have adopted here is very different. In our approach, scenes are represented holistically, as a single entity that does not need to be parsed further into distinct objects. This viewpoint renders both problems (scene and object recognition) to have equivalent complexity.

It has been shown how the holistic scene representation strategy can be used for classifying outdoor scenes and also for incorporating contextual influences on the task of object recognition. In this paper we have focused on indoor scene recognition. This choice of domain allows us to test the versatility of the representation scheme and also enables us to build a system that has important potential applications. For instance, a mobile robot intended for use in houses, factories or offices will benefit from an ability to identify its surroundings visually. Additionally, since our scheme is experimentally motivated and uses neurally plausible computational mechanisms, it can serve as a model of place recognition by biological systems. Indeed, a straightforward extension of this system can serve to model the 'place cells' that have been reported in the hippocampal tissue of rats and other animals [McNaughton et al., 1996].

There are several interesting directions in which to extend this work. For instance, even though the current implementation of our system does not require individual object recognition as a prerequisite for scene recognition, the two processes can be made to operate synergistically. Thus, inferences about scene identity can prime an object detection system and the latter's results can, in turn, improve scene classification performance. Another interesting direction involves determining whether the training from one environment can be useful for making inferences in a novel setting. In other words, can a system trained in the office space at MIT use some of its knowledge to infer place categories in a different office? Finally, there is the issue of exploring other representation strategies. We have experimented with one particular choice of representation scheme. There may well be other strategies that can robustly encode the stable and discriminant aspects of scenes with less computational expense.

References

[1] Borenstein, J., Everett, H.R., and Feng, L. (1996). 'Where am I?' Sensors and Methods for Mobile Robot Positioning. Technical Report, The University of Michigan. http://www.eecs.umich.edu/ johannb/pos96rep.pdf

[2] Cartwright, B. A. and Collett, T. S. (1987). Landmark maps for honeybees. Biological Cybernetics, 57, 85-93.

[3] Carson, C., Belongie, S., Greenspan, H., and Malik, J. 1997. Region-based image querying. *Proc. IEEE W. on Content-Based Access of Image and Video Libraries*, pp: 42–49.

[4] Clarkson, B., Mase, K., and Pentland, A. 2000. Recognizing user's context from wereable sensor's: baseline system. Vismod Technical Report 519.

[5] De Bonet, J. S., and Viola, P. 1997. Structure driven image database retrieval. *Adv. in Neural Information Processing* **10**.

[6] Gorkani, M. M., and Picard, R. W. 1994. Texture orientation for sorting photos "at a glance". *Proc. Int. Conf. Pat. Rec.*, Jerusalem, Vol. I, 459–464.

[7] Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision*, 20(11):1254–1259.

[8] Janzen, D. H. (1971). Euglossine bees as long-distance pollinators of tropical plants. Science, 171, 203-205.

[9] Lipson, P., Grimson, E., and Sinha, P. 1997. Configuration based scene classification and image indexing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Puerto Rico, pp 1007-1013.

[10] McNaughton, B.L., Barnes, C.A., Gerrard, J.L., Gothard, K., Jung, M.W., Kniermim, J.J., Kudrimoti, H., Qin, Y., Skaggs, W.E., Suster, M., and Weaver, K.L. 1996. Diciphering the hippocampal polyglot: the hippocampus as a path integration system. *J. Exp. Biology* 199:173-185.

[11] Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial Envelope. *International Journal of Computer Vision*. 42(3): 145–175.

[12] Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., and Ballard, D.H. 1996. Modeling saccadic targeting in visual search. NIPS'95. MIT press.

[13] Schiele, B., and Crowley, J.L. 1997. Recognition without Correspondence using Multidimensional Receptive Field Histograms. M.I.T. Media Laboratory, Perceptual Computing Section Technical Report No. 453

[14] Schyns, P. G., and Oliva, A. 1994. From blobs to boundary edges: evidence for time- and spatial-scale dependent scene recognition. *Psychological Science*. 5:195-200

[15] Sirovich, L., and Kirby, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4, 519-524

[16] Szummer, M., and Picard, R. W. Indoor-outdoor image classification. In *IEEE intl. workshop on Content-based Access of Image and Video Databases*, 1998.

[17] Torralba, A., and Oliva, A. 1999. Scene organization using discriminant structural templates. *Proc. Of Int. Conf in Comp. Vision*, ICCV99, 1253-1258.

[18] Torralba, A., and Oliva, A. Depth perception from familiar structure. submitted.

[19] Torralba, A., and Sinha, P. (2001). Statistical context priming for object detection. *IEEE Proc. Of Int. Conf in Comp. Vision.*

[20] Vailaya, A., Jain, A., and Zhang, H. J. 1998. On image classification: city images vs. landscapes. *Pattern Recognition*, 31:1921–1935

[21] Wehner, R., and Raber, F. (1979). Visual spatial memory in desert ants, Cataglyphis bicolor (Hymenoptera Formicidae). Experientia, 35, 1569-1571.

[22] Wehner, R. (1987). Spatial organization of foraging behavior in individually searching desert ants, Cataglyphis (Sahara desert) and Ocymyrmex (Namib desert). In Pasteels, J. M., and Denebourg, J. L., eds., From Individual to Collective Behavior in Social Insects. Basel: Birkhauser, 15-42.