Visual Learning for Landmark Recognition

Yutaka Takeuchi, Patrick Gros,
Martial Hebert, Katsushi Ikeuchi
School of Computer Science, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh PA 15213
hebert@cs.cmu.edu, ki@cs.cmu.edu
http://www.cs.cmu.edu/~{hebert,ki}

Abstract[1]

Recognizing landmark is a critical task for mobile robots. Landmarks are used for robot positioning, and for building maps of unknown environments. In this context, the traditional recognition techniques based on strong geometric models cannot be used. Rather, models of landmarks must be built from observations using image-based visual learning techniques. Beyond its application to mobile robot navigation, this approach addresses the more general problem of identifying groups of images with common attributes in sequences of images. We show that, with the appropriate domain constraints and image descriptions, this can be done using efficient algorithms as follows: Starting with a "training" sequence of images, we identify groups of images corresponding to distinctive landmarks. Each group is described by a set of feature distributions. At run-time, the observed images are compared with the sets of models in order to recognize the landmarks in the input stream.

1. Introduction

Recognizing landmarks in sequences of images is a challenging problem for a number of reasons. First of all, the appearance of any given landmark varies substantially from on observation to the next. In addition to variation due to different aspects, illumination change, external clutter, and changing geometry of the imaging devices are other factors affecting the variability of the observed landmarks. Finally, it is typically difficult to use accurate 3D information in landmark recognition applications. For those reasons, it is not possible to use many of the object recognition techniques based on strong geometric models.

1. This research was sponsored in part by Office of Naval Research (ONR) under Contract N00014-95-1-0591 and, in part, by DARPA under contract DAAE07-96-C-X075 monitored by TACOM. Patrick Gros was supported by a NATO fellowship. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ONR, DARPA or the U.S. Government.

The alternative is to use image-based techniques in which landmarks are represented by collection of images which are supposed to capture the "typical" appearance of the object. The information most relevant to recognition is extracted from the collection of raw images and used as the model for recognition. This process is often referred to as "visual learning".

Progress has been made recently in developing such approaches. For example, in object modeling [Gros et al.], 2D or 3D model of objects are built for recognition applications. An object model is built by extracting features from a collection of observations. The most significant features are extracted for the entire set and are used in the model representation. Extension to generic object recognition were presented recently [Carlsson, 1996].

Other recent approaches use the images directly to extract a small set of characteristic images of the objects which are compared with observed views at recognition time. For example, the eigen-images techniques are based on this idea.

Those approaches are typically used for building models of single object observed in isolation. In the case of landmark recognition for navigation, there is no practical way to isolate the object in order to build models. Worse, it is often not known in advance which of the objects observed in the environment would constitute good landmarks.

A similar problem, although in a different context, is encountered in image indexing, where the main problem is to store and organize images to facilitate their retrieval [Lamiroy et al., 1996] [Schmid et al., 1996]. The emphasis in this case is the kind of features used and the type of requests that can be made by the user.

Our approach tries to combine these two categories of systems. In a training stage, the system is given a set of images in sequence. The aim of the training is to organize these images into groups based on similarity of feature distributions between images. The size of the groups obtained may be defined by the user, or by the system itself. In the latter case, the system tries to finds the most relevant groups, taking the global distribution of the images into account. In a second step, the system is given new images, which it tries to classify as one of the learned groups, or in the category of unrecognized images.

The basic representation is based on distributions of different feature characteristics. All these different kinds of histograms are computed for the whole image and for a set of sub-images. Tests are used to compare these histograms and define a distance between images. This distance is then used to cluster the images into groups. Each group is then characterized by a set of feature histograms. When new images are given to the system, the algorithm evaluates a distance between these images and the groups. The system determines to which group this image is the closest, and a set of thresholds is used to decide if the image belongs to this group.

The main goal of the work presented here was to explore the use of tools and methods in the field of image retrieval to the problem of landmark recognition. It is clear that the global architecture of the system is close to that of object recognition systems [Gros et al.]: a training stage in which 3D shape, 2D aspects, or groups, are characterized is followed by a recognition stage in which this information is used to recognize the models, objects or groups in new images. The difference comes from the wide diversity of the images and from the groups which are not reduced to a single aspect of an object. The two challenging tasks which we concentrate on in the remainder of the paper are to define these groups more precisely as sets of images, and to build automatically a characterization for each group.

The first section of the paper deals with the feature distributions, their computation and their comparison. The second section addresses the computation and the characterization of groups. The third section concerns the classification of new images according to the groups previously defined. Experimental data and results are presented in the fourth section.

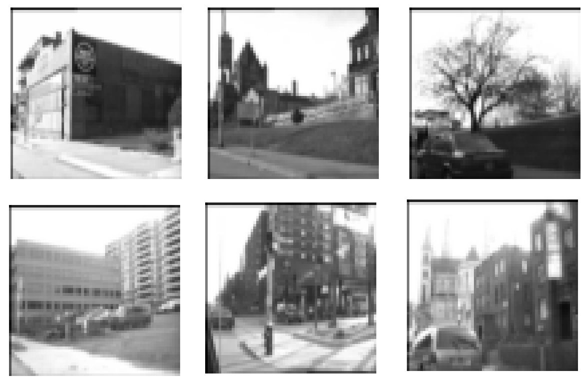

Figure 1: Sample set of images from a typical training sequence. The complete training sequence contains 176 images taken over approximately 800 meters.

2. Representing Images

In this section, we give a brief overview of the features used for representing images. Since an individual feature can be characteristic of an aspect of an object, but probably fails to characterize well a set of aspects, we use a statistical description of a large number of features as our basic representation Representing Feature Distributions

Five basic features are currently used: distributions of normalized color and intensity, edges, segments, and parallel segments. Additional features can be added as needed. The basic image representation is a set of feature distributions. Edges are computed using Deriche's edge detector [Deriche, 1987]. Segments are computed as a polygonal approximation of the edges [Horaud et al., 1990].

Among the characteristics computed from these five features are: color and grey levels; edge density, orientation, length and position; segment density, orientation, length, and position; parallel segment density, orientation, length, and position, and, finally, the angles between adjacent segments.

Feature densities are computed in two ways: first as the ratio of the number of features (edges, segments and parallel segments, respectively) to the area of the image or sub-image concerned; second as the ratio of the sum of all the feature lengths to the area of the image or sub-image. All the other measurements are computed and the results stored in histograms. Each bucket of the length histograms indicates how many features a length has in a given interval.

The position of a particular object in an image may vary substantially between observations. Therefore, it is important to build the representation in a way that allows for different placements of the object with similar resulting feature distributions. This is done by subdividing the image into smaller chunks, in which the feature distributions are computed. All these histograms and densities, except those relative to feature position, are computed for the whole image and for the sub-images obtained by dividing the image by 4, then 9, and then 16. The position histograms are computed only for the global image, i.e., 90 densities and 333 histograms are computed to characterize each image.

2.1. Comparing Feature Distributions

The feature distributions from two images are compared using a distance similar to the chi-square distance. This distance, in its simplest form, evaluates the probability that two sets of data, here the histograms or the densities, are derived from the same theoretical distribution. If the distributions are $h = (h_i)$ and $l = (l_i)$, their difference is computed as [Press et al., 1992]: $\mathrm{d}(h, l) = \Sigma_i(h_i - l_i)^2/(h_i + l_i)$.

The main problem is to derive a global distance between two images from the individual distances computed for each type of density and histogram. The distance d_{ij} between two images i and j is defined as the linear combination of the distances between individual feature distributions: $d_{ij} = \Sigma_k \lambda_k \mathrm{d}(h, l)$, where the sum is taken over all the feature distributions used for representing the images.

When nothing is known about the distributions and their range of variation, all the weights can be taken equal to one. This simple approach gives good results in practice, but a better approach is to compute the weights based on the relative scales of the feature distributions. For each kind of density or histogram, the distance between every pair of images is computed, and the variance σ^2_k of these distances is derived. The coefficient relative to this particular distribution can be chosen as $\lambda_k = 1/\sigma^2_k$. This choice of weights has the effect of normalizing the distributions and of assigning the same relative importance to all the partial distances used.

3. From Images to Landmarks

Given a sequence of images, we now want distinctive landmarks, that is, we want to split the sequence into groups of images, and find a characterization of each of these groups which allows further classification.

This step is difficult to do fully automatically in general. The main reason is that there is not a task-independent definition of the type of image groups that are needed. Our approach is to use task constraints to guide the grouping process. Specifically, given an initial grouping of images, we select groups based on three constraints. First, only the groups that contain a large enough number of images from different aspects are retained. Second, groups that do not provide significant discrimination are discarded. This is important to ensure that, at recognition time, only the groups that can be easily distinguished are used as models. Finally, the recorded sensor position for each training image is used for ensuring that the groups are spatially coherent.

3.1. Computing Initial Image Groups

Once the distance matrix is computed, a simple agglomerative method is used to split the image set into initial groups. First each image is put in a different group. Then the two closest groups are grouped and the distance matrix is updated. Finally, the algorithm iterates the previous step until an ending condition is verified.

Let $|L|$ denote the number of elements of the image group L. The update of the matrix consists of suppressing the two lines and two columns i and j corresponding to the groups I and J which are grouped, and then adding a new line and column for the new group formed. The diagonal term of the new line added is:

$$\frac{|I|(|I|-1)d_{ii} + |J|(|J|-1)d_{jj} + 2|I||J|d_{ij}}{|I|(|I|-1) + |J|(|J|-1) + 2|I||J|} \qquad (1)$$

The non diagonal term corresponding to the new group and a group k is:

$$\frac{|I|}{|I \cup J|}d_{ik} + \frac{|J|}{|I \cup J|}d_{jk} \qquad (2)$$

These formulas show that, at each iteration, the only information needed is the distance matrix and the number of elements in each group. Each term of the matrix is thus the mean distance between the images of two groups, or the mean distance of the images of a same group.

3.2. Controlling the Grouping Algorithm

The grouping algorithm described above is general. In particular, it does not incorporate a control structure that stops the grouping process when groups of images corresponding to recognizable landmark are formed. An automatic method was developed for controlling image grouping.

Given a set of image groups, let us denote the distance matrix by (d_{ij}) and the number of images of the group i by n_i. If the images used to learn the groups form a representative sample, and if they are spread nearly uniformly in representation space, the probability that an unknown image will be classified in the group i (p_i) or in no group at all (p_0) can be evaluated. If n denotes the number of groups:

$$p_i = \frac{d_{ii}}{\sum_j d_{jj} + \frac{2}{n-1}\sum_{j \neq k} d_{jk}} \qquad p_0 = \frac{\frac{2}{n-1}\sum_{j \neq k} d_{jk}}{\sum_j d_{jj} + \frac{2}{n-1}\sum_{j \neq k} d_{jk}} \qquad (3)$$

These formulas state that the probability p_i that a new image belongs to a group is proportional to the size d_{ii} of this group, and that the probability p_0 of being in no group is proportional to the distances d_{jk} between the groups. The factor $2/(n\text{-}1)$ is used to compensate the number of non-diagonal terms of the distance matrix with respect to the number of diagonal terms.

In the case where the images are known not to have a uniform distribution in a region of their representation space, this can be taken into account by using these other formulas:

$$p_i = \frac{n_i d_{ii}}{\sum_j (n_j+1)d_{jj} + \frac{2}{n-1}\sum_{j \neq k} d_{jk}}$$
$$p_0 = \frac{\frac{2}{n-1}\sum_{j \neq k} d_{jk}}{\sum_j (n_j+1)d_{jj} + \frac{2}{n-1}\sum_{j \neq k} d_{jk}} \qquad (4)$$

In these new formulas, not only the size of the groups is taken into account but also their density, which is proportional to n_i. The probability p_0 is also a function of the size of the groups: for the same distances between the groups, the smaller the groups, the bigger their density, and the smaller is the probability of having a new image between them.

The evaluation function is the entropy associated with this set of probabilities $S = -\Sigma_i\, p_i \ln p_i$, and the process is stopped when this entropy is maximal. This maximizes the information provided to the user by each classification request.

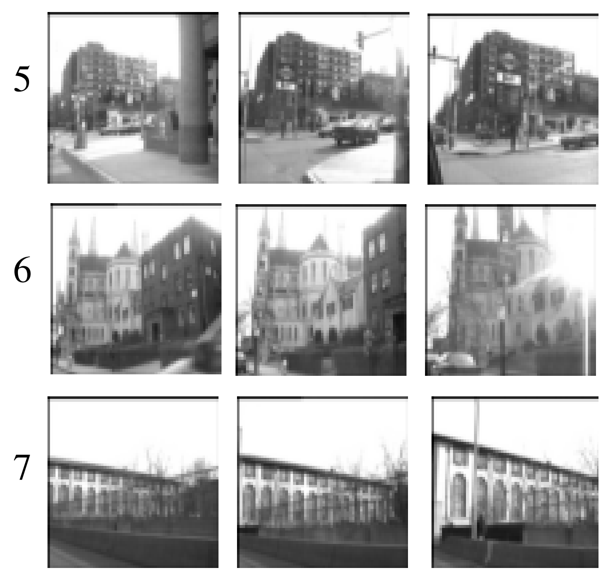

Figure 2: Three of the groups extracted from the training sequence of Figure 1. Only three images are shown for each group.

3.3. Using Domain Constraints

An important constraint in the context of landmark recognition is that the images are ordered in the training sequence. In fact, the position of the vehicle is recorded for each image. Using this information, grouping is limited to images for which the vehicle positions were close to each other, thus ensuring spatial consistency.

In general, given an image, it would be necessary to consider other images for grouping in a radius around the corresponding vehicle position. In practice, images in the training sequence are digitized at approximately equal intervals. As a result, it sufficient to consider for grouping with image i only images j such that $|i - j| < r$, where r is the maximum extent of observability of any given landmark. This constraint reduces the computational complexity of the grouping algorithm, and it guarantees that the image groups correspond to spatially coherent landmarks.

4. Recognition

Given a set of image groups C_i, characterized by their mean vector v_i, their eigenvalues $\lambda_{i1}...\lambda_{il}$, and their eigenvectors $v_i^1 ... v_i^l$, the problem is to compare these groups to a new image whose characteristic vector is v. The eigenvalues are positive, and the eigenvectors of a group are a family of orthonormal vectors, and $v - v_i$ may be decomposed with respect to this family: $v - v_i = \Sigma_j a_j v_i^j + r$. The distance between v and the group C_i is derived as:

$$d(C_i, v) = \sqrt{\sum_j \frac{a_j^2}{\lambda_i^j + 1} + \|r\|^2} \qquad (5)$$

This formula allows us to find which is the closest group to an image. The problem is then to decide if the image really belongs to this group. We again use the task constraints. Intuitively, consecutive images should be classified in a consistent manner. Since we know the spatial extent from which a particular group of images is visible in the training sequence, we can eliminate the inconsistent classification results as unreliable.

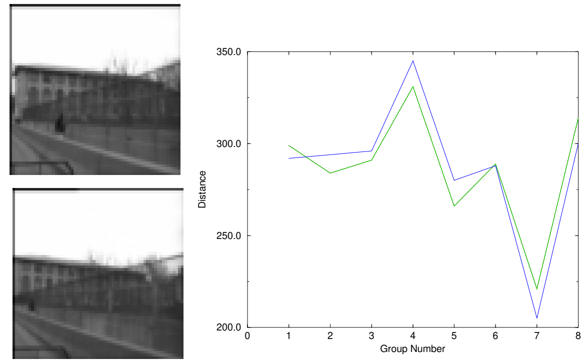

Figure 3: Comparison of two new images (top) with the groups shown in Figure 2. The distance graph of the left image is shown as a solid line. The images are correctly classified in group 7.

5. Examples

The recognition algorithm was tested using several sequences taken from a moving vehicle. Figure 4 shows sample images from a test sequence taken over the same course as the training sequence of Figure 1. A total of 68 images were digitized from the test sequence at regular intervals. The test set of images was segmented manually into subsets corresponding to the landmarks identified in the training sequence. This provided the "ground truth" for evaluating the performance of the algorithm. Vehicle position was recorded using the INS system onboard a HMMWV used in a separate Unmanned Ground Vehicle (UGV) project [Hebert et al., 1996].

Using the algorithm outlined above, all the images in the test sequence were compared with the landmarks found in the training sequence. The images corresponding to a distinct distance minimum are labeled with the corresponding landmark number. Images that do not correspond to a distance minimum for any landmark are left unclassified. As we noted earlier, our goal is not to label every image but rather to correctly recognize the ones corresponding to the most salient landmarks.

Figure 5 illustrates the classification algorithm for three different landmarks. The graphs show that, for those three landmarks, the distance minimum is attained at the correct images in the test sequence. Figure 7 shows a view of all the landmarks recognized in the path travelled in the test sequence.

Figure 4: Sample images from a test sequence. The complete sequence is 68 images over a course similar to the one used in Figure 1 for the training sequence. The orientation of the camera, and the illumination characteristics are substantially different from those of the training sequence.

Four landmarks are recognized in this example. Figure 7 also illustrates the potential use of landmark recognition for position estimation.

Two types of errors may occur during recognition. First, images that should match a landmark are matched to a different landmark. We call those images misclassified images. Second, images that should match a given landmark are not matched with any landmark. We call this second class of images unclassified images. To reduce the error rate, we use the sequential constraint described earlier. This constraint is quite effective in practice. Figure 6 shows the error statistics for the recognition algorithm with and without sequential constraint.

All the examples given so far involve images taken in urban environments. We have also conducted experiments in natural environments by collecting training sequences, extracting groups of distinctive images, and recognizing them in test sequences. Figure 8 shows two example groups computed from a typical training sequence. The error rate, is comparable to the one obtained in urban scenarios.

6. Conclusion

Results on image sequences in real environment show that visual learning techniques can be used for building image-based models suitable for recognition of landmarks in complex scenes. The

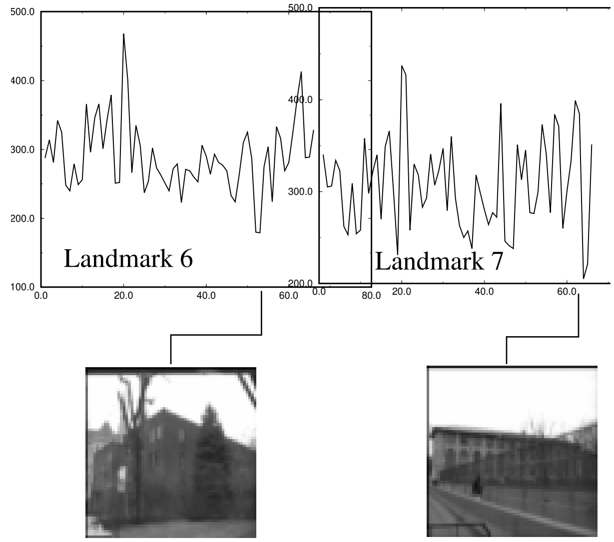

Figure 5: The classification algorithm illustrated on two landmarks. Each graph shows the distance between all the images of the test sequence of Figure 4 the groups found in the training sequence (Figure 2.) The graphs are shown for landmarks 2 and 7. The graphs show that the distance is minimum for the correct landmark.

Urban Environment:

	distance only	**with sequential constraint**
correct	72%	93%
misclassified	19%	0%
unclassified	9%	7%

Natural Environment:

	distance only	**with sequential constraint**
correct	84%	97%
misclassified	6%	0%
unclassified	10%	3%

Figure 6: Performance of the recognition algorithm on the two example sequences. Images are labeled as “misclassified” if they are matched to the wrong group; they are labeled as unclassified if they belong to a group but are not matched.

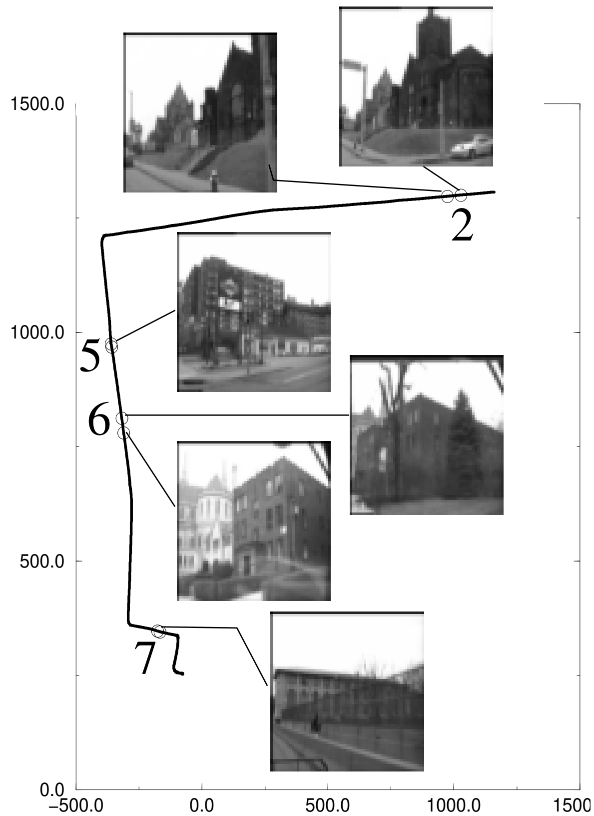

Figure 7: Overhead view of the path followed while collecting the images of the test sequence of Figure 4 (distances are indicated in meters.) Four landmarks are correctly identified. Example images from the test sequence are shown for each landmark.

Figure 8: Images of one of the landmarks found in a sequence of images in a natural environment.

approach performs well, even in the presence of significant photometric and geometric variations, provided that the appropriate domain constraints are used. In the case of mobile robot navigation, domain constraints include the sequential nature of the images, and the discriminability of landmarks.

Our goal is to demonstrate the use of landmark recognition for navigation. Specifically, we will show that rough position estimation and navigation based on the relative positions of landmarks can be achieved using image-based landmark recognition. Several limitations of the approach need to be addressed. First of all, rejection of unreliable groups need to be improved. In particular, the selection of the parameters controlling the grouping need to be implemented in a principled manner. Second, images that do not contribute information should be filtered out of the training sequences.

References

[Carlsson, 1996] S. Carlsson. Combinatorial Geometry for Shape Representation and Indexing. *Proc. International Workshop on Object Representation for ComputerVision*. Cambridge, England, April 1996.

[Deriche, 1987] R. Deriche *Using Canny's Criteria to Derive a Recursively Implemented Optimal Edge Detector*. Int. Journal of Computer Vision 1(2), pages 167--187, 1987

[Gros et al.] P. Gros, O. Bournez and E. Boyer. *Using Local Planar Geometric Invariants to Match and Model Images of Line Segments*. To appear in Int. J. of Computer Vision and Image Understanding.

[Hebert et al., 1996] M. Hebert, C. Thorpe, A. Stentz. *Intelligent Unmanned Ground Vehicle.s*. Kluwer Publishers. 1996.

[Horaud et al., 1990] R. Horaud, T. Skordas and F. Veillon. Finding Geometric and Relational Structures in An Image *Proc. of the 4th European Conf. on Computer Vision,* Antibes, France April 1990

[Lamiroy et al., 1996] B.Lamiroy and P.Gros. Rapid Object Indexing and Recognition Using Enhanced Geometric Hashing. *Proc. of the 4th European Conf. on Computer Vision*, Cambridge, England, pages 59--70, vol. 1, April 1996.

[Press et al., 1992] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. *Numerical Recipes in C*. Second Edition, Cambridge University Press 1992.

[Schmid et al., 1996] C. Schmid and R. Mohr. Combining Greyvalue Invariants with Local Constraints for Object Recognition. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Francisco, California, USA. pages 872--877, June 1996