Robust Place Recognition using Local Appearance based Methods

Gregory Dudek and Deeptiman Jugessur

Centre for Intelligent Machines
3480 University Street
McGill University
Montréal, Québec, Canada H3A 2A7

Abstract

We present an approach to the automatic recognition of locations or landmarks using single camera images. Our approach is to learn visual features in the appearance domain that can be used to characterize an object or a location. These features are defined statistically and then are recognized using principal components in the frequency domain.

We show that this technique can be used to recognize specific objects on varying backgrounds, as well as environmental features.

1 Introduction

In this paper we present a formulation of coarse robot position recognition based on learned features. We focus on the recognition of locations or specific objects, but not on computing the precise quantitative position of either the robot or the object with respect to the robot. We presume that quantitative position estimation problems can, however, be addressed using the types of computations we present here.

The features we used can be derived from almost arbitrary pose-dependent sensor data, although in this paper we consider only the use of camera data. Our recognition technique can be applied to two types of pose estimation scenario: recognition of an entire view of a scene (for example a room), or recognition of a specific class of objects that can be deposited in the scene to act as landmarks. Our objective is thus to be able to recognize a familiar learned scene or object given a test image. This would allow a mobile robot to recognize its current location. Several methods permit a precise position estimate to be computed given an approximate one using landmarks from either vision or sonar[1, 2, 3]

Our approach is based on learning visual features that characterize either an object in the scene or a view of a scene (such as a room) in an off-line learning stage. Then, during on-line execution, a mobile robot can recognize these known objects. (In the interests of succinctness, we will hereafter refer to both familiar items as well as familiar views of a scene as "objects").

The key contribution of this paper over preceding work is with respect to the robustness of our approach to rotation, partial occlusion and other variations in the scene content. Known objects are recognized from a set of characteristics appearance-based features. Each of these features is computed to limit its sensitivity of background image content, as well as rotation and translation.

2 Background

Place recognition (or *localization*) methods can be broadly classified into local and global methods. Local methods are those that produce a quantitative pose estimate assuming the approximate neighborhood is known, while global methods are those which operate over a large environment and often produce only a qualitative estimate. The approach to position estimation presented in this paper is a global position estimation scheme.

Several authors have considered the use of linear subspace methods, often referred to as principal components analysis or *eigen-* methods to recognize objects [4] or compute robot pose [5, 6]. In their basic form, these methods represent images that containing objects of interest in a low dimensional subspace. The distance of a test image from known sample images then is used to compute its identity. Such "appearance based" methods have met with considerable success in applications such as face recognition, but since they use the entire image they are sensitive to occlusion, rotation, illumination variation and scale changes.

Sim and Dudek [7, 8] used PCA methods on small sub-windows of a larger image to compute quantitatively accurate local pose estimates from image samples in an environment. Their approach assumes that a very approximate pose estimate is available as input (i.e. what room the robot is in) and achieves a numerically accurate pose estimate by combining information from multiple observations using probabilistic weighting of measurements.

In recent work, Lowe[9] has also proposed recognizing objects using small image samples. In contrast to this paper, Lowe appears to use large numbers of samples each with a low disambiguating power. He relies on votes techniques for recognition. Similarly, Schmid et al. [10] has considered object recognition using sub-windows extracted using the Harris operator each of which makes only a small contribution to the final identification. Our work, in contrast, uses a small number of measurements each of which has substantial disambiguating power.

3 Problem Statement

Scene and landmark recognition with computer vision using PCA methods has been limited by the need to use essentially global image data. In particular, PCA-based methods have been sensitive to variations in the background behind objects of interest, and to occlusion or change in parts of the scene. Traditional global approaches fail to recognize objects successfully if more than some 1/3 of the image changes (and sensitivity is often much worse than this). Our work sets out to accomplish appearance-based object recognition while remaining robust to variations in the background, changes in sub-parts of the scene, or occlusion of a substantial fraction of the image. In addition, we seek a recognition system that exhibits rotation invariance since our robots often take images while their tilt is unpredictable.

4 Approach

4.1 Principal Component Analysis Overview

To perform the actual classifications of the images to be recognized ,an image compression technique known as principal component analysis (PCA) is used. This allows images to be compared in a lower dimensional space (lower than the number of pixels N in an image) by computing the eigenvectors of the covariance matrix $\mathbf{Q}$ of the training image set (the training image set being the set of recognizable objects). These eigenvectors form an orthogonal basis set for representing individual images in the set. Images to be recognized are projected onto this eigenspace and matches are made by examining the Euclidean distance between points in this space.

Various methods exist to compute the eigenvectors of $\mathbf{Q}$ and we choose the singular values decomposition of the matrix $\mathbf{P}^T$ where $\mathbf{P}^T$ is the transpose of the matrix $\mathbf{P}$ where each row consists of a training image from which the average of all the training images has been subtracted.

Since any M by N matrix $\mathbf{A}$ ($M \geq N$) can be written as

$$\mathbf{A} = \mathbf{U}\mathbf{X}\mathbf{V}^T \tag{1}$$

we chose $\mathbf{A}$ to be $\mathbf{P}^T$. $\mathbf{U}$ and $\mathbf{V}$ are M by N and N by N matrices, respectively, with orthonormal columns, and $\mathbf{X}$ is an N by N diagonal matrix containing the *singular values* of $\mathbf{A}$ along its diagonal [11]. Specific to our needs is the fact that the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{P}\mathbf{P}^T$ and the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{P}^T\mathbf{P}$, hence the eigenvectors of the covariance matrix are obtained.

4.2 Using attention operators and sub-windows to make PCA robust

Since each row of $\mathbf{P}$ contains the intensity values of an entire image consisting of a recognizable object with no preprocessing, classic PCA as outlined above is very sensitive to translations, rotations (planar or non-planar), scaling of the object within the image and occlusions. Furthermore as no a priori segmentation can be done in the image to be recognized, backgrounds which differ from those within the training set result in misclassification of the objects to be recognized as only raw intensity values are considered. This is due to the fact that all the images are compared in the eigenspace constructed by $\mathbf{P}$. Objects to be recognized which are off center, rotated (planar or non-planar), scaled, occluded even partially or on different backgrounds relative to those within $\mathbf{P}$, when projected onto the eigenspace result in points that are not necessarily close to their corresponding training image eigen points. We account for some of these problems by the introduction of an interest operator which chooses points within the images. Sub-windows are cropped around the chosen points and instead of performing PCA on the entire image, it is performed on these sub-windows. We use a symmetry based context free attention operator [12] which is independent of segmentation.

Our recognition algorithm consists of a training phase and a testing phase. During the training phase we run the interest operator on the set of images which we want to recognize, crop around these interest points and build $\mathbf{P}$. In the testing phase, we run the interest operator on the image to be recognized, process the information in the sub-windows obtained to account for planar rotations and varying backgrounds and project them onto the eigenspace.

In the absence of noise the attention operator will choose the same points of interest in testing images as those it chose during the training phase (which is mostly the case for the operator we use). This achieves translation invariance as all that matters is the image data in the immediate neighborhood of the attention point.

Multiple interest points are chosen for the recognition of an object. So long as a sufficient fraction of the interest points associated with the object are recovered, the object can be recognized. Since a voting scheme is devised for all the sub-windows around the interest points chosen, see section 4.5, partial occlusions cause the interest points that are chosen on the occluding object in the image to cast erroneous votes. The points chosen on the object itself, cast good votes and can thus at times (depending on the degree of occlusion) reliably identify the object itself.

4.3 Feature Locality

The image content around each selected interest point is extracted in a manner that is not purely local. Emphasis should be given to the immediate neighborhood of the interest point chosen while image data as one heads to the periphery of the sub-window should hold less weight. Multiplying the data within the sub-window with a two dimensional Gaussian reduces sensitivity to distal points which may be on the background.

4.4 Rotation Invariance

The use of a Fourier basis for the sub-windows provides rotation invariance. This depends on representing the data in polar form and on the shift theorem. One of the properties of the two-dimensional Fourier Transform is the *shift theorem*. Given a function $f(x,y)$ in the spatial domain, its Fourier Transform gives us a function $F(u,v)$ in the frequency domain. The *shift theorem* states that the Fourier Transform of $f(x-a, y-b)$, where a and b are constants, is:

$$e^{j2\pi(au+bv)}F(u,v) \tag{2}$$

This property can be exploited in our situation to help achieve planar rotation invariance around the chosen

(a) (b) (c) (d)

Figure 1: Polar Sampling of an Ice tea can. (a) and (b) are (x,y) images with (b) rotated by 45 degrees. (c) and (d) are their corresponding polar (θ,r) images. Note how a 45 degree rotation in (x,y) image turns into a shift in θ (θ,r) image

interest points. This is due to the fact that both $f(x,y)$ and $f(x-a,y-b)$ have the same same amplitude spectrum in the frequency domain as:

$$|e^{j2\pi(au+bv)}F(u,v)| = |F(u,v)| \tag{3}$$

Given a sub-window in the image, we first convert it into a polar coordinate system. Thus any rotations about the interest point are reflected as shifts in θ in the polar image, see Figure 1. Once a polar representation of the sub-window is obtained, it is multiplied by a two-dimensional Gaussian with a standard deviation of a quarter the length of the square sub-window. This is effectively the Hamming window as a notion of continuity is introduced between the two ends of each line of the image and thus one gets rid of some of the potential high frequency components in the spectrum introduced.

We then proceed to obtain the amplitude image through the Fourier transform of the two dimensional polar image outlined above. This amplitude image is invariant to any rotations about the center of the window and thus we achieve planar rotation invariance. Note that rotations about arbitrary points are handled automatically since for any interest point they can be described as a translation and a rotation about the center.

4.5 Classification

In the off-line training phase, the set of data gathered from all the sub-windows of all the training images are collectively used to create the database of recognizable objects or images. Application of PCA to this allows for the construction of a sub-space suitable for recognition.

On-line recognition is performed by associating the interest regions from a test image with their training image counterparts. This is achieved by successively projecting each sub-window onto the eigenspace created off-line and finding the closest known eigenpoint corresponding to the most similar training sub-window image.

A voting mechanism is added as multiple interest points represent an image or object to be recognized. The following algorithm is used to accumulate the data obtained by the projection of all the interest points for a given test image:

- For each interest point x in the test image
  - Project x onto the eigenspace to get the eigenpoint $\tilde{X}$
  - Find the closest projected training point $\tilde{Y}$ in the eigenspace to $\tilde{X}$
  - Find $D = dist(\tilde{X}, \tilde{Y})$ where $dist$ is Euclidean distance.
  - Given $\tilde{Y}$ find its corresponding training image T
  - Add the value of $1/(D+\epsilon)$ to T's weight W. Note that ϵ is a constant.

The training image with the largest value of W is the closest match to the test image.

5 Experimental Results

Tests were conducted on two sets of colour images to exemplify the recognition of both objects that act as landmarks, as well as entire views. We used a database of 20 objects (Figure 2) taken from the Columbia Object Image Library (COIL) and a database of 20 scenes (Figure 8) taken at various locations in our lab.

Recognition performance on these two training sets was evaluated using sub-window sizes of 10 by 10 pixels together varying numbers of interest points per image. In this paper we show results using 64 interest points per image, but similar findings have also been obtained with as few as 5 interest points extracted per image. A complete analysis of the results obtained from varying these parameters is part of our ongoing research.

Figure 3 shows four examples of the test images used for object recognition. Note that the objects within the images are rotated, partially occluded and placed on artificial backgrounds. Figures 4, 5, 6 and 7 show the values of W as outlined in the algorithm of section 4.5 for the images shown in figure 3. The ordering of the test images corresponds to the ordering of the results. The fourth test case is particularly interesting in that it illustrates how the approach can occasionally fail: it shows erroneous results caused by the fact that the majority of the interest points chosen fall on the background. Interestingly enough, the closest match found is the clay mug which has a similar

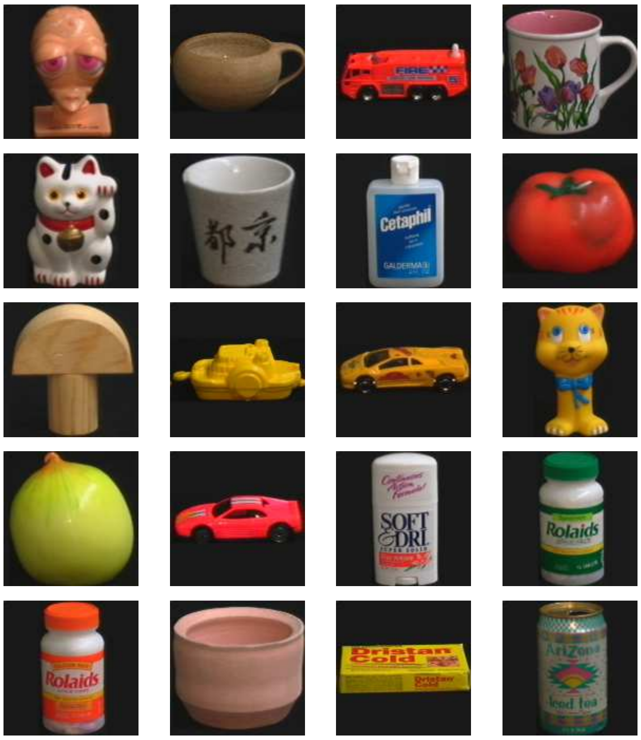

Figure 2: Database of training objects (all 128x128 images)

Figure 3: Test images used for obtaining results of Figures 4, 5, 6 and 7

colour to the background in the failing test case. An issue this raises is that the attention operator must be suitably tuned with respect to both the scale and structure of the types of objects of interest. While this is a subject of ongoing research; we note that this type of artifact can also be seen with human observers (the "Where's Waldo" series of books serves as a familiar example).

Figure 9 shows four of the test images used for the room recognition. Note that in these test cases, the camera is no longer aligned in the fashion it was when the training pictures were taken. The big black triangle in the first image is to simulate the occlusion of half the test image. Figure 10 shows the training images chosen by our system which best represent to test cases shown in figure 9.

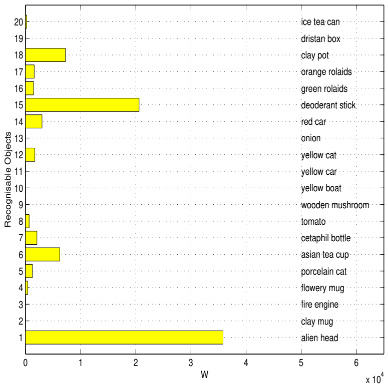

Figure 4: Recognition of the alien head from Figure 3

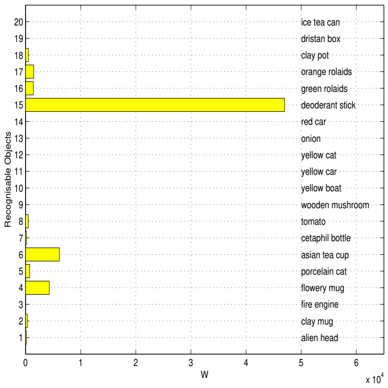

Figure 5: Recognition of the deoderant stick from Figure 3

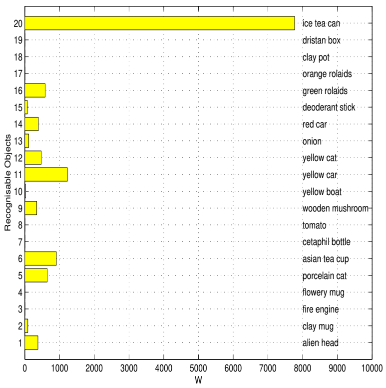

Figure 6: Recognition of the ice tea can from Figure 3

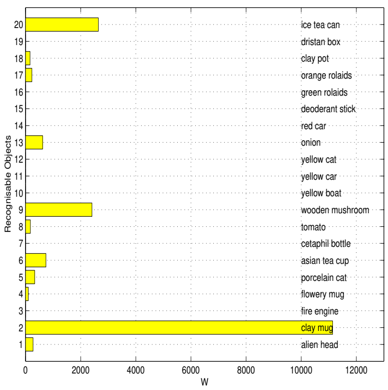

Figure 7: Failed recognition of the last ice tea can Figure 3

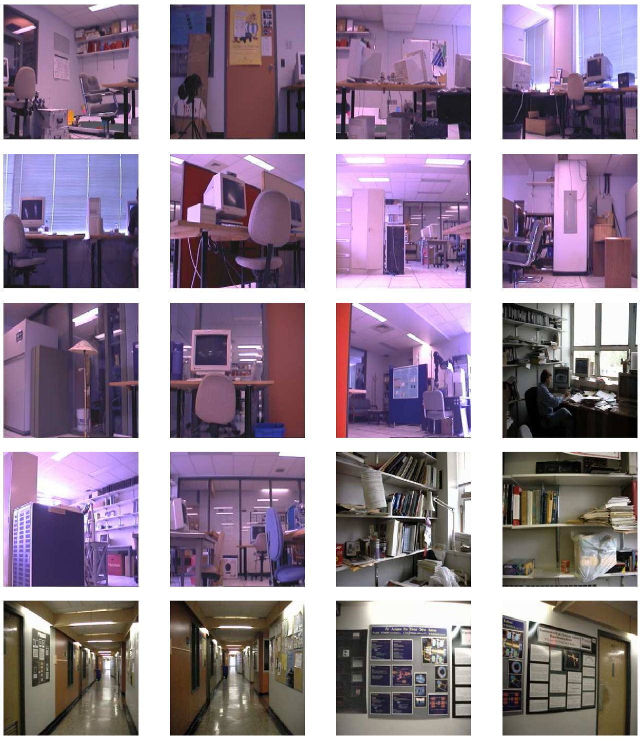

Figure 8: Database of training rooms (all 256x256 images)

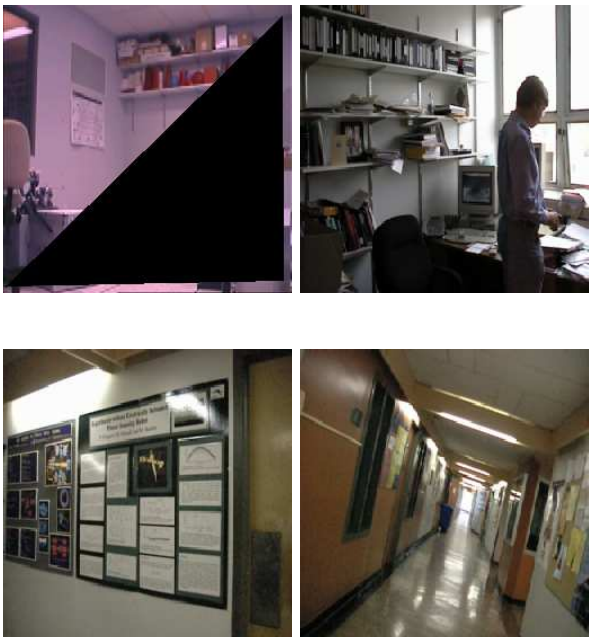

Figure 9: Test images used for obtaining results of Figures 10

Figure 10: Training images from figure 8 chosen by our system when using Figures 9 as test cases.

6 Conclusions

In this paper we have presented an approach to vision based recognition for navigation. In particular, an approach to the recognition of landmark objects and of scenes that would allow places in the world to be recognized by a mobile robot. Subsequent to this recognition, various existing approaches for precise localization could be used (for example to compute the specific pose with a room of interest) [7].

Our approach to recognition is based on appearance-based subspace projection, although by using an attention mechanism we manage to achieve robustness in the fact of background variations (for landmarks), large amounts of occlusion, planar rotation, and other distortions.

Unresolved issues relate to the use of multiple attention operators, which might allow for even greater robustness in the face of very complex occluding objects or extremely large degrees of occlusion. We are also examining the question of how to automatically select an appropriate number of interest points from an image to assure confident yet efficient recognition. We note, in passing, that since only small image subwindows are used for the recognition process, it is highly efficient as compared to traditional full-image PCA methods.

References

[1] M. Betke and L. Gurvits, "Mobile robot localization using landmarks," *IEEE Trans. on Robotics and Automation*, vol. 13, pp. 251–263, April 1997.

[2] R. Sim and G. Dudek, "Learning visual landmarks for pose estimation," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, (Detroit MI), IEEE Press, May 1999.

[3] K. Sugihara, "Some location problems for robot navigation using a single camera," *Computer Vision, Graphics, and Image Processing*, vol. 42, pp. 112–129, 1988.

[4] M. Turk and S. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[5] S. K. Nayar, H. Murase, and S. A. Nene, "Learning, positioning, and tracking visual appearance," in *Proc. IEEE Conference of Robotics and Automation*, (San Diego, CA), pp. 3237–3244, IEEE Press, May 1994.

[6] F. Pourraz and J.L.Crowley, "Continuity properties of the appearance manifold for mobile robot position estimation," in *Proc. IEEE Workshop on Perception for Mobile Agents*, (Ft. Collins, CO), IEEE Press, June 1999.

[7] R. Sim and G. Dudek, "Learning environmental features for pose estimation," in *Proc. IEEE Workshop on Perception for Mobile Agents*, (Ft. Collins, CO), IEEE Press, June 1999.

[8] R. Sim and G. Dudek, "Robot positioning using learned landmarks," *McGill TR CIM-98-1170*, June 1998.

[9] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. International Conference on Computer Vision*, (Corfu, Greece), IEEE Press, Sept. 1999.

[10] C. Schmid, "A structured probabilistic model for recognition," in *Proc. IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, (Fort Collins, CO), pp. 485–490, IEEE Press, June 1999.

[11] *Digital Image Processing*, ch. 10. Prentice-Hall, 1996.

[12] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context free attentional operators: the generalized symmetry transform," *International Journal Of Computer Vision*, vol. 14, pp. 119–130, 1995.