Local Appearance for Robust Object Recognition

Deeptiman Jugessur and Gregory Dudek

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

Abstract

We present an approach to appearance-based object recognition using single camera images. Our approach is based on using an attention mechanism to obtain visual features that are generic, robust and informative. The features themselves are recognized using principal components in the frequency domain.

In this paper we show how the visual characteristics of only a small number of such features can be used for appearance-based object recognition that is not confounded by planar rotations or background clutter.

1 Introduction

In this paper we consider appearance-based object recognition with robustness that is increased over that exhibited by traditional methods. In particular, we outline two improvements to standard PCA-based recognition that increase its resistance to background variation, in-plane rotation and partial occlusion.

Object modeling based on linear subspace approximation of intensity images has developed into a powerful technique for object recognition. This is traditionally accomplished by a principal components analysis of an ensemble of training images, a long standing image processing technique reintroduced to the vision community by Turk and Pentland [1] and by Murase and Nayar [2] among others. Work in this field has been constantly hindered by two well-known problems: the use of global images for PCA-based recognition leads to sensitivity of the recognition of a foreground object to background content; the technique is very sensitive to any rotation of the foreground object, even simple planar rotations.

In this paper, we address both of these issues simultaneously and, in addition, perform recognition efficiently. Our approach is based on using an attention mechanism to obtain visual features that are generic, robust and informative. These features then serve as the cues to object recognition. The features themselves are recognized using principal components in the frequency domain.

The use of local features extracted by an attention operator allows us to develop an approach to recognition that can extract discriminative cues from an object of interest even though much of the image may contain irrelevant content, or when the object to be recognized is occluded. This achieves partial insensitivity to background clutter, although it does entail a dependence on the availability of one or more appropriate attention operators. The truly adventurous might take this as a justification of the wide variety of alternative attention mechanisms observed in the human visual system [3]. The use of amplitude spectra to subsequently recognize the features that attract attention allows the features to be recognized independent of their rotation in the plane. This, in turn, allows object recognition to be achieved independent of 2D rotations.

In this paper we describe the function of our system using a single attention operator based on symmetry [4]. This operator has a number of attractive features including scale invariance, biological relevance, and stability. Our approach has also been tested with alternative attention operators although a discussion of the associated issues are outside the scope of this paper. Suffice it to say that in an applied context, we would propose to use the technique described here with features extracted by multiple attention subsystems.

The outline of this paper is as follows. In Section 2 we briefly consider some relevant background research, followed by our problem statement in Section 3. In Section 4 we provide the details of the methods we used to implement our object recognition system. This in turn is followed by some of our experimental results (shown in Section 5). Finally, in Section 6 we summarize our observations and discuss some further issues and ongoing work.

2 Background

Several authors have considered the use of linear subspace methods, often referred to as principal components analysis or *eigen-* methods to recognize objects [5, 1] or compute robot pose [2]. In their basic form, these methods represent images that contain 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 appli-

cations such as face recognition, but since they use the entire image the are sensitive to occlusion, rotation, illumination variation and scale changes.

In recent work, Lowe[6] has also proposed recognizing objects using small image samples. In contrast to this paper, Lowe initially generates a large number of samples, of which only a few are necessary for recognition. He relies on voting techniques for recognition. Similarly, Schmid et al. [7, 8] has considered object recognition using small windows extracted using the Harris operator each of which makes only a small contribution to the final identification. Our work, in contrast, uses a smaller number of measurements each of which has substantial disambiguating power. Kohtaro Ohba and Katsushi Ikeuchi [9], present independently developed work of a very similar flavour as our own. They too choose to perform the recognition of their extracted features using PCA, however they only concern themselves with the recognition of occluded objects.

3 Problem Statement

Global PCA-based methods have been sensitive to variations in the background behind objects of interest; the location of the object to be recognized within the image; changes in the orientation of the object and to occlusion or changes 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 some rotation invariance (specifically planar rotation invariance) since our robots often take images while their tilt is unpredictable.

4 Approach

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. The smaller the distance between the point representing the image to be recognized and another point (one of the projected training set images), the better the match. Dimensionality reduction comes into play as it can be shown that despite the fact that all N eigenvectors are needed to represent the images exactly, only a small number k ($k \ll N$) is generally sufficient for capturing the primary appearance characteristics of the recognizable objects [10]. These k eigenvectors correspond to the k largest eigenvalues of the covariance matrix $\mathbf{Q}$. Comparisons are thus made in this lower dimensional eigenspace.

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.

4.1 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 [4] 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 a selected group of these interest points (see Section 4.2) 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.

4.2 Filtering Interest Points

Once an interest map is obtained from the attention operator, interest points are chosen from the map such that the information content within the sub-windows around the interest point is plentiful. The degree of information within a sub-window is determined by computing the standard deviation of intensity values within that window. A standard deviation threshold is computed and any cropped sub-windows with standard deviations less than that threshold are rejected.

Another level of discrimination is introduced by requiring that sub-windows cropped around chosen interest have a minimum degree of overlap with other chosen sub-windows. This results in interest points being chosen over a large portion of the object to be recognized instead of multiple windows being chosen around a few points which are considered most interesting by the operator.

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. This also reduces the sensitivity to the shape of the window. Advantages are also gained for the achievement of rotation invariance as outlined in the next section.

4.4 Rotation Invariance

The use of a Fourier basis for the sub-windows chosen around the interest points provides rotation invariance. One of the properties of the two-dimensional Fourier

(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

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:

$$e^{j2\pi(au+bv)}F(u, v) \tag{1}$$

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

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

Given a sub-window in the Cartesian coordinate system, we first convert it into a polar coordinate system by sampling in a circular fashion with increasing radii from the center of the window. 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 due to the discontinuities in the discrete two dimensional image which we treat as a continuous signal. Section 4.3 also outlines advantages of such an operation.

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 kd-tree was implemented to search the eigenspace efficiently for the nearest neighbor of the projected eigenpoint. Since such a data structure also allows one to retrieve the n nearest neighbors efficiently, part of our ongoing research involves selecting not just the nearest neighbor but choosing an eigenpoint from the n nearest neighbors.

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.

ϵ is a constant which is introduced to account for outliers which can cast large votes. All our experiments in Section 5 use an ϵ value of one.

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

5 Experimental Results

In this section, we report the results of object-recognition tests using a database of assorted objects. Since we were unable to find a suitable database (such as Columbia University's COIL database) that included sufficient background variation, we were forced to construct of own database[1] Figure 2 shows a sample of the set of objects with which the database of images was created. Three views were taken for each object; figure 2 only shows one such view per object for six of the 14 recognizable objects used for our experiments. [2]

Recognition performance on the training set was evaluated using sub-window sizes of 10x10 and 20x20 pixels respectively. The eigenspace itself was created using 50 interest points per training image and the 30 most significant eigenvectors were used for classification. Tests were conducted to evaluate the accuracy of the recognition as a function of the number of interest points. Figure 3 shows 6 of the 17 test samples which were used to obtain the results shown in Figure 4. Note that these test images include training objects which are placed on various non-uniform backgrounds. Test images also include examples where the object to be recognized undergoes limited non-planar rotations. Despite that, recognition was still reliably achieved. Note also that some of the objects are partially occluded.

Figure 4 shows two performance plots. The first graph shows the percentage of images that were recognized plotted against the number of interest points used as input. Note that with 50 interest points all of the objects shown in Figure 4 are recognized with a sub-window size of 10x10. The second graph shows an increase in the reliability of the recognition as one increases the number of interest points. The reliability is simply defined as the difference in vote strength between the two most preferred models, more specifically the average of the sum of all the differences of W (as outlined in Section 4.5) between the recognized object and the second best candidate object. It is interesting to note that roughly 70 percent of the test images were recognizable with only 5 interest points. The area spanned by 5 interest points with a 10x10 window size corresponds to 500 pixels which is merely 0.0016 of a 640 by 480 pixel image . This exemplifies the difference between our approach to attention-based appearance modeling and others based on the use of far more numerous but "weaker" features.

An analysis of how the chosen window sizes (10x10 and 20x20 in the results shown here) affects recognition is currently part of our ongoing research. From the results we see that while there seems to be no significant change in the percent of images recognized vs. number of interest points plot, the reliability of the recognition was better when we used the larger 20x20 sub-window. We also found that a lot of the images which were not recognized even with 50 interest points when

[1] These images are available to other researchers at http://www.cim.mcgill.edu/~dudek/objects

[2] Note that all these color images are actually of size 640 by 480 while the images shown in this paper have been cropped and scaled appropriately for legibility. The test images shown have not been cropped but simply re-scaled for display.

Figure 2: 6 of the 14 training objects used for the database of recognizable objects. (all 640x480 images)

Figure 3: 6 of the 17 test images (all 640x480 images) used for results shown in figure 4

using 10x10 sub-windows, were successfully recognized with the larger sub-window, see Figure 6 for an example of such an image. This is not surprising as the information content per sub-window is greatly increased and hence recognition is more reliable.

Figure 5 is 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. 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. In the cases shown, the background exhibits symmetric properties which attracts the attention of the interest operator used as the scale over which it was working was not appropriately tuned. The images show where the 50 interest points (10x10 sub-windows) were chosen and one can clearly see that the majority of these do not lie within the object of interest.

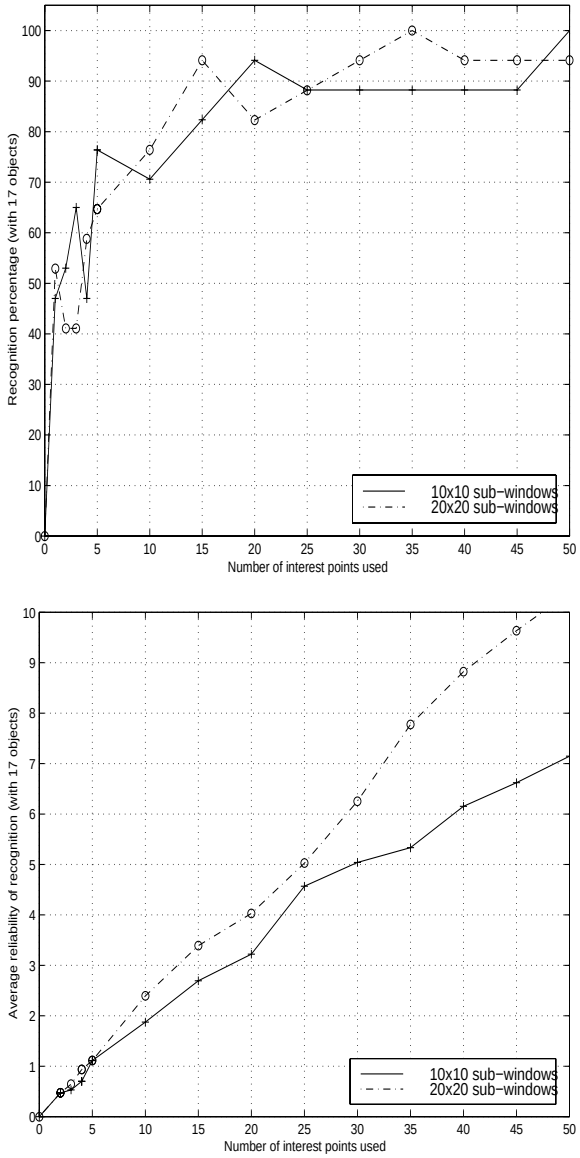

Figure 4: Results obtained when varying the number of interest points on images such as those shown in figure 3. The first graph shows the recognition performance on all the 17 images (of which 6 are shown in figure 3) as a percentage. The second graph shows the reliability of the recognition performed. See Section 5 for a definition of reliability.

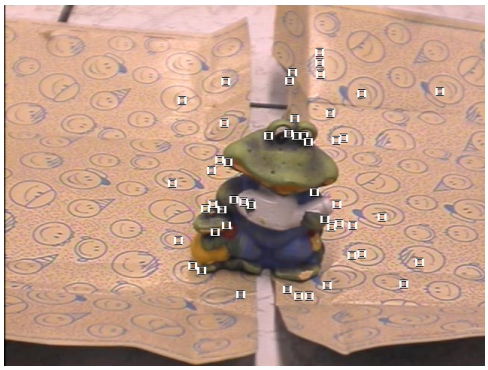

Figure 5: Example test image (containing sub-windows around the interest points) which caused a failure (all 640x480 images).

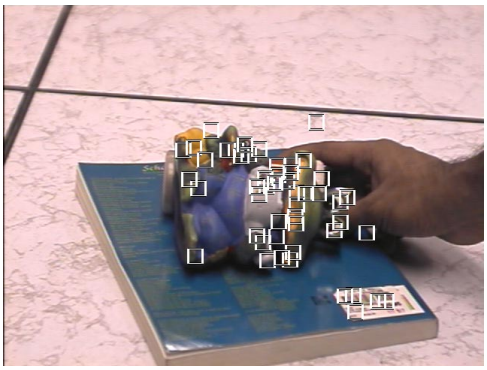

Figure 6: Example of a test image which was reliably recognized using 50 interest points and a window size of 20x20 but failed for the 10x10 case

6 Conclusions

In this paper we have presented a refinement of appearance-based object recognition using principal components analysis. Rather than directly using a subspace of the entire image, we propose the use of a set of cues selected with an attention operator to drive the recognition process. The regions about these cues are then learned and recognized using principal components analysis in the Fourier domain. By using the frequency domain to characterize the appearance of each attention point, we achieve invariance to two dimensional rotations. The fact that we use a consensus of a number of small attention points to identify an object makes the technique robust and efficient. This robustness is achieved since we only need a limited fraction of the interesting points on an object to be visible. The efficiency results from the fact that the set of local features we use comprises only a small fraction of the entire image (although we must process the entire image with the attention operator). In principle, the approach should also accomodate scaling fairly easily. In other work we have also considered the effects of alternative attention operators but none so far have improved on the results presented here.

In this paper we have considered the performance of the approach as a function of the number of interest points used and briefly discussed the impact of the size of the subwindows extracted around each attention point. Over our ensemble, "large" windows (0.13 percent of the entire image) provide more reliable results (even with only 5 feature points) but with smaller windows we can still achieve good recognition rates by using larger numbers of features. Clearly, the attention operator must be well suited to the ensemble of objects being recognized but our method is independent of the specific operator used. In ongoing work, we are examining the use of multiple alternative operators in conjunction with one another.

References

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

[2] 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.

[3] A. Triesman, "Perceptual grouping and attention in visual search for features and objects," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, no. 2, pp. 194–214, 1982.

[4] 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.

[5] Duda and Hart, *Pattern Classification and Scene Analysis*. New York. NY: Wiley, 1973.

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

[7] 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.

[8] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–534, 1997.

[9] K. Ohba and K. Ikeuchi, "Detectability, uniqueness, and reliability of eigen-windows for robust recognition of partially occluded object s," *IEEE Pattern Analysis and Machine Intelligence*, vol. 19, no. 9, pp. 1043–1048, 1997.

[10] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.