ROBUST OBJECT RECOGNITION USING LOCAL APPEARANCE BASED METHODS

Deeptiman Jugessur

Department of Computer Science

McGill University, Montréal

9 October 2000

A Thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of
Master of Science

ABSTRACT

In this thesis we present an approach to appearance-based object recognition using single camera images. The 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. We show that we can perform robust appearance based object recognition by using the visual characteristics of only a small number of such features. The technique is robust to planar translations and rotations of the object being recognized due to our polar sampling in the frequency domain. We are able to recognize objects on different types of background clutter due to a masking technique we've developed. We are also able to handle a degree of occlusion as we make use of multiple features for the purposes of recognition.

The same approach is further applied in the field of robotics to provide a means for the automatic recognition of locations or landmarks in scenes typically encountered by mobile robots. Hence instead of only recognizing objects, we also present a means of using the same computational model to recognize locations, thus performing coarse localization.

RÉSUMÉ

Dans cette thèse nous présentons une approche basée sur l'apparence pour reconnâitre des objects en utilisant une seule image de caméra. Notre approche est basée sur un méchanisme d'attention pour obtenir des caractéristiques visuelles qui sont génériques, robustes et significatives. Les caractéristiques elles-mêmes sont reconnues en utilisant les composantes principales dans le domaine des fréquences pour tenir compte des rotations planaires de l'objet analysé. Nous montrons qu'on peut reconnâitre des objets par seulement utiliser une petite quantité de ces caractéristiques visuelles. Notre travial a pour objectif d'accomplir une reconaissance d'objet basée sur l'apparence, tout en demeurant robuste aux variations d'arrière-plan (nous avons dévéloper une technique pour ça qu'on appel "masking"), aux changements de sous-ensembles de la scène, ou aux occlusions d'une fraction substantielle de l'image (parce qu'on utilise plusieurs de caractéristiques visuelles par image). De plus, nous proposons un système de reconnaisance qui démontre une certaine invariance aux rotations (plus spécifiquements, une invariance aux rotations planaires) étant donné que nos robots font souvent des acquisitions d'images alors que leur inclinaisons est imprévisible.

La même approche est aussi utiliser pour une application de robot mobile où nous fournissons la capacité de reconnaissance automatique de la position de points d'intérêt dans une scène typique, et aussi pour la reconnaissance de la scène entière.

ACKNOWLEDGMENTS

First and foremost I'd like to thank my gifted supervisor, Gregory Dudek, who never ceases to amaze me with the amount of tasks he can simultaneously juggle while making everyone of his students feel as though their work, truly is *the* most important.

Of the many friends who helped me I'd like to reserve a special thank you for: Richard Unger, who's always so full of inspirational ideas; Robert Sim, for being my coding guru; Eric Bourque, for teaching me the importance of pedantic code; Sylvain Bouix, for keeping me grounded with countless discussions; Sandra Polifroni, who's always willing to help, and Vida Dujmović, who's thoughts, friendship and support will always be priceless.

I'd also like to thank everyone in the lab, who at some point or another have been involved in my thesis, especially François Bélair. Finally I'd like to thank my parents for supporting me throughout my masters but more accurately, my entire life.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

1.1. Motivation

Our sense of sight is arguably the most important of our five senses. The benefits of a human's ability to recognize what he or she sees, require no explanations. In computer vision, attempts to replicate such an ability in machines is plagued with problems due largely to the fact that we are unsure as to how such tasks are performed in our brains. Simple tasks which we take for granted, for example, the ability to recognize a chair in a photograph of a furnished room or the ability to recognize a friend's face in a crowd of faces are unsolved computational problems in object recognition.

One approach to image recognition, is to compare an input image to many other images that have been previously stored. Various methods have been proposed to implement the comparison between the input image to be recognized and the database of images which the recognition system can exploit. While little is known about the exact representations and techniques used by the human mind when performing object recognition, it is clear that the overall appearance of the object plays a critical role in its perception. Recent physiological findings suggest that the human visual system represents objects by a set of two-dimensional views rather than (or in addition to) a single object-centered three dimensional model [**14**]. The technique used in this thesis allows a system to recognize three-dimensional objects from their appearance in two-dimensional images. Hence instead of creating complicated three-dimensional models of the object to be recognized (which usually requires manual input of information into the vision system) and comparing such computer aided design (CAD) models to CAD models in the database of recognizable objects, simple two-dimensional snapshots of the object to be recognized are all that is required by the recognition system.

To perform the actual comparisons between these two-dimensional images an image compression technique known as principal component analysis (PCA) (also known as the Karhunen-Loeve transform [**8**]) is sometimes used. This allows images to be compared in a lower dimensional space (lower than the number of pixels in an image) by computing the eigenvectors of an image set. 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 distances 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 images), the better the match.

Since classical PCA based recognition (global PCA) requires no preprocessing of the image (the image is considered in its entirety), it is very sensitive to variations in lighting, scale, translation, rotation (planar and non-planar), occlusion and the background.

The focus of this thesis is to make global PCA less sensitive to the background, planar translations and rotation and occlusion. Unlike global PCA we propose a methodology whereby we fuse an attention mechanism into the system and only use PCA on the processed intensity information around a local neighborhood of interest points chosen by the attention mechanism.

1.2. Problem statement

In this thesis we consider appearance-based object recognition with robustness that is increased over that exhibited by traditional methods. In particular, we outline improvements to prior appearance-based recognition that increase its resistance to background variation, in-plane translation and rotation and partial occlusion. Furthermore we use the same framework for the automatic recognition of locations or landmarks in large-scale environments (for example an office like environment) using single camera images.

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 pattern classification [**7**] and image processing technique reintroduced to the vision community by Turk and Pentland [**29**] and by Murase and Nayar [**14**] among others. Work in this field has been hindered by two well-known problems:

(i) The use of global images for PCA-based recognition leads to sensitivity of the recognition of a foreground object to background content.

(ii) The technique is very sensitive to any rotation of the foreground object, even simple planar rotations.

In this thesis, 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 in color 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 [**27**]. 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 thesis we describe the function of our system using three attention operators. The bulk of the results presented however are based on a radial symmetry operator developed by Yeshurun et al. [**18**] (See Section 3.8). 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 based on edge density [**3, 4**], edge orientation [**3, 4**] and the KLT operator [**26**] which bases attention points on track-ability (see Section 3.8). Suffice it to say that in a particular application, we would propose to use the technique described here with features extracted by *multiple* attention subsystems.

1.3. Outline of our solution

We perform object recognition using a simple *training*, *testing* paradigm defined by *off-line* and *on-line* phases. In the *off-line* training phase, we run an interest operator on our set of images containing the objects we wish to distinguish. This process determines areas within the training images which are considered salient and thus suitable for recognition. Sub-windows are extracted around these salient features and are further processed to provide planar rotation invariance. This involves the use of a two dimensional Fourier transform to obtain a rotation invariant amplitude spectrum from the information within the sub-windows. An eigenspace is then constructed using this information from a large set of sub-windows from many objects for the purposes of PCA-based classification.

The *on-line* testing phase once again consists of running the same interest operator on a single *test* image. The underlying principle for successful classification of the object within the test image lies in the fact that a large fraction of the salient points returned by the interest operator coincide with those from the training image counterpart (ie. the training image containing the object under inspection in the test image). Classifications themselves are performed by projecting the processed test image sub-windows onto the eigenspace created in the training phase and finding the nearest neighbors of the sub-windows. The Euclidean distance between the corresponding points in the eigenspace is a measure of the similarity of the sub-window images in the $L2$ norm [**14**]

1.4. Outline of this thesis

The outline of this thesis is as follows. In Chapter 2 we briefly consider some relevant background research in the field of object recognition with an emphasis being placed on previous work in PCA-based object recognition. In Chapter 3 we provide the details of the principles used to develop our object recognition system. Within this chapter, we also present the interest operators we used for our experiments in Section 3.8. We continue with our experiments and their corresponding results in Chapter 4 where we show the performance of our system in the context for object recognition (see Section 4.1) and place recognition (see Section 4.2). Finally, in Chapter 5 we summarize our observations and discuss some further issues and ongoing work.

CHAPTER 2

Background and previous work

2.1. PCA based object recognition

Principal component analysis (PCA) is a standard, well-known technique used widely in many fields, primarily those concerned with multivariate analysis. The central idea behind PCA is to reduce the dimensionality of the data set, while retaining as much as possible of the variation present in it. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated (orthogonal), and which are ordered such that the first *few* retain most of the variation present in *all* the original variables. The actual computation of the principal components reduces to eigenvalue/eigenvector problem for the covariance matrix (see Section 3.2) [**10**].

PCA was first introduced by Pearson in 1901 and was also independently developed by Hotelling in 1933 [**10**]. Pearson gave the earliest description of what we now call PCA when he implemented a methodology to find lines and planes which best fitted a set of points in p-dimensional space. Hotelling's approach was based in the ideas of *factor analysis* (often confused with PCA, in part due to Hotelling's 1933 paper where principal components were introduced in the context of providing a small number of 'more fundamental' variables which determine the values of the p original values). Both PCA and factor analysis aim to reduce the dimensionality of a data set but the approaches used to do so are different. Hotelling's motivation was to seek a smaller fundamental set of independent variables which determine the values' of the original p variables. He introduced the term *components* to avoid confusion with other uses of the word factor in mathematics. He chose his *components* so as to maximize their successive contributions to the total of the variances of the original variables. The components which he thus derived were called principal components.

PCA has been extensively used in many fields ranging from meteorology to economics. Dimensionality reduction is often performed in the analysis of multivariate data, irrespective of the field where the data is being analyzed. Applications of such a dimensionality reduction scheme range from the visualization of multi-dimensional data by examining their first two or three principal components, to data compression for efficient storage purposes.

Despite being a standard, established technique in mathematics and pattern recognition [**7, 8**], it was not until 1991 that it gained popularity in the machine vision field for image recognition. Matthew Turk and Alex Pentland (1991) [**29**] used PCA for face recognition and their approach proved to be very successful[1]. They credit Sirovich and Kirby [**25**] for their own motivation to use *eigenfaces* (their term for the principal components representing their set of recognizable faces) as Sirovich and Kirby used PCA for image compression, specifically the compression of face images. Turk and Pentland present a near-real time face recognition system that not only uses PCA for the face recognition but also to detect faces within an image. To account for varying backgrounds they multiply the face image with a two dimensional Gaussian, thus accentuating the center of the face while placing less of an emphasis on the background and as they state, varying hair-styles in people. While their results are excellent, problems arise as they assume faces to be facing the camera and upright (hence they have no planar rotation invariance) and the existence of shadows within the images are not accounted for as the entire face is examined irrespective of illumination models. Variations in lighting direction and facial expressions pose

[1]While Turk and Pentland are often credited with reintroducing PCA to the vision community it should be noted that Murase *et al.* [**9**] applied PCA earlier than Turk and Pentland in the computer vision field for handwritten character recognition problems.

significant problems for the classic PCA model for face recognition as the images are examined as a whole. As the data deals entirely with the raw intensity values of the pixels within an image, slight changes in the lighting direction for example can lead to gross misclassifications in the eigenspace. Belhumeur *et al.* [**1**] account for such variations in lighting direction and changes in facial expressions by using Fisher's Linear Discriminant, which while similar to the PCA or eigenface techniques proposed by Turk and Pentland, has the advantage of better handling the changes stated above. They present some of the disadvantages of the eigenface model from a clustering perspective and present improvements in what they refer to as the *fisherface* model.

Murase and Nayar [**14**] use PCA, applying it to the general object recognition problem as well as for pose estimation. The authors detail the advantages of being able to recognize three dimensional objects through simple two-dimensional intensity images, such as not having the overhead required to construct computer aided design (CAD) models of the objects of interest.

Their recognizable object database (training database) consists of sets of images of specific objects where for each object a number of images are stored under varying lighting conditions and poses. Poses are varied by rotating the objects through known angles about their principal axes. A *universal eigenspace* (the eigenspace computed for *all* the objects) and a set of *object eigenspaces* (eigenspaces computed for *each* object) are computed and maintained for recognition and pose recovery. The former is used to recognize an incoming test object and the latter to compute its pose. The paper details the constructions of these *parametric* eigenspaces (parametric as they are dependent on the parameters of lighting and pose variations) and also the assumptions made when these eigenspaces are computed. The variations in lighting and angle are performed such that the projected points of the images form a smoothly varying *manifold* in the eigenspace. Since consecutive images are strongly correlated, their projections in eigenspace are close to one another, hence one can hope to obtain such a smooth manifold under suitably constrained circumstances. If the manifolds of two objects intersect in the universal eigenspace, the intersection corresponds to poses of the two objects whose images are very similar in appearance. They claim that such images are inherently ambiguous as they simply do not contain sufficient information for the unique identification of the object under scrutiny. Given a test image, it is first projected onto the universal eigenspace and the nearest object manifold is found. This results in the identification of that object. Once identified, the object is projected onto the appropriate manifold and the nearest eigenpoint is found. Such an eigenpoint represents the image with a known pose and lighting condition which best matches that of the test candidate.

Murase and Nayar's results convincingly show the success rates and advantages of such an appearance based methodology. However they observe that all their experiments are dependent on the use of constrained environments where the background of the objects under scrutiny are always simple and well known (typically black). They state the problems which arise due to the segmentation problem and also note the systems inability to handle occlusions. Planar rotations of the objects are not considered or even addressed.

A problem of substantial interest and relevance to robotics is the visual positioning and tracking problem. Nayar *et al.* [**15**] apply the above stated methods in such a context. They detail the construction involved to allow for the automatic positioning of a robot manipulator at a desired coordinate using only two dimensional intensity images. The visual positioning problem can be stated as the automatic positioning of a robot manipulator at a desired position with respect to an object, given an input image of the object taken at the manipulators current position. The positioning problem is solved by finding the eigenpoint on the object manifold that provides the minimum distance d between the eigenpoint representing the projected image taken at the manipulators current position, and the manifold of the object. The visual tracking problem is simply solved by repeated iterations of the positioning problem.

In more recent work in pose estimation using visual information, Sim and Dudek [**24, 23**] used similar PCA methods on small sub-windows of a larger image (instead of using the entire image) to compute quantitatively accurate local pose estimates from image samples in an environment. Their approach assumed that a very approximate pose estimate was available as input (i.e. what room the robot was in) and achieved a numerically accurate pose estimate by combining information from multiple observations using a probabilistic weighting of measurements. While the use of sub-windows achieved robustness in the face of partial occlusions for robot pose estimation, and reduced sensitivity to illumination variations, that approach remained sensitive to rotation. If it were to be used for object recognition, it would also have exhibited sensitivity to background changes.

2.2. Current relevant work in object recognition

In recent work, Lowe[**11**] has also proposed recognizing objects using small image samples. The image samples chosen, or the features, are selected and processed such that they are invariant to image scaling, translation, and rotation. Furthermore they are partially invariant to illumination changes and affine or 3D projections. The paper describes how such features are generated and selected in a method called the Scale Invariant Feature Transform (SIFT). SIFT keys are generated for each selected feature vector and the keys themselves are used as input to a nearest-neighbor indexing method that identifies candidate object matches.

Lowe obtains the features using a staged filtering approach. Key locations are identified in scale space by the search for locations that are maxima and minima of a difference-of-Gaussian function. This is done efficiently by the construction of an image pyramid with re-sampling between each level. Each level of the pyramid is generated by repeatedly applying bilinear interpolations for re-sampling to Gaussian smoothed images. Maxima and minima of this scale space function are determined by comparing each pixel in the pyramid to its neighbors. The SIFT keys thus generated are then further processed for stability and each key location is also assigned a canonical orientation, thus making the image descriptors rotation invariant.

While providing an excellent approach with very convincing results, Lowe's experiments leave the reader with many questions. The results provided show examples of successful recognition under quite chaotic conditions. Lowe's ability to recognize objects when the background is highly cluttered is quite remarkable and he also demonstrates his systems ability to handle occlusions as only a subset of the generated features in a test case need to be matched to training features for robust recognition. However all his results are demonstrated with only three objects and no analysis is presented on the number of features that are generally required for robust recognition.

Another recent and notable image recognition approach is offered by Schmid [**20**]. Once again image features are extracted for the purposes of recognition. Her paper derives a probabilistic model for recognition based on local descriptors (the extracted features) and the spatial relations between the descriptors. The classifications of query images is performed through the correspondences of local descriptors found in the database of recognizable images with those found in the query image. They construct 9-dimensional vectors for each descriptor such that the vectors contain intensity invariants which describe the local neighborhood of the interest points as well as local intensity derivatives. The vector invariants are also computed at several scales to allow for a certain degree of scale invariance. Correspondences are determined by comparing these descriptor vectors together with their geometrically constrained spatial relations to one another.

While the paper is more geared towards image retrieval from a database than object recognition, she presents a robust solution to the problem and her ideas can be extended to the object recognition problem. Schmid, Mohr and Bauckhage provide an in depth study of interest operators in [**21**] and as a direct result of that study, they choose to use the Harris corner and edge detector [**19, 20**]. The operator has the advantage of providing the most *repeatable* results under varying conditions from the operators they examine [**21**]. They define repeatability as the percentage of the total observed interest points which are repeated between two images of the same image where the second image has undergone transformations such as image rotation and scaling.

In [**20**], Schmid provides a probabilistic model for the scheme proposed in [**19**] and she convincingly shows the advantages of using a Bayesian framework over the voting mechanism proposed in the older paper. The results demonstrated [**19, 20**] are excellent, however, the images which are used as test samples do not vary significantly from those which are within the database. Since the images themselves are to be identified as opposed to objects within the images, the authors do not concern themselves with the background within the images.

Finally we consider an independently developed method similar to our own which is presented by Kohtaro Ohba and Katsushi Ikeuchi [**16**]. While the authors concern themselves mainly with the recognition of partially occluded objects, their approach consists once again of using extracted features from the image. Furthermore they propose an *eigenwindow* approach whereby an eigenspace is constructed using these extracted features. Query objects are recognized through a voting scheme developed on the basis of Euclidean distances between projected query image *eigenwindows* and their nearest neighbor eigenwindows representing the database images. The eigenwindows themselves are chosen using the track-ability criteria proposed in [**22**] and are further processed for the purposes of robustness and information content. As their approach is primarily concerned with the recognition of occluded objects, they do not address the issues of rotation and scale invariance of the selected eigenwindows. The background content in the images is also not addressed.

2.3. Our contributions

This thesis contributes a powerful approach to appearance-based recognition and goes on to provide a detailed analysis of the key performance parameters. While several authors attempt to solve the object recognition problem using feature selection, they fail to simultaneously recognize the objects on textured backgrounds, under rotations and translations of the object in the view field and under cases where the object is partially occluded from view. Our approach successfully achieves these task by combining an automated means of selecting features using appropriately tuned interest operators and the use of PCA in the frequency domain for classification. Working in the frequency domain (detailed in Section 3.10) allows for planar rotations of the object under scrutiny as well as translations in the plane. Furthermore we provide a means of recognizing objects independent of a wide variety of backgrounds they appear on through our masking technique (detailed in Section 3.11). We handle a large degree of occlusion as long as a significant percent of the features chosen fall on the occluded object. We also contribute to the field of mobile robotics by providing a novel application of our object recognition technique for the automatic recognition of locations or landmarks in typical scenes encountered by mobile robots.

CHAPTER 3

Methods

In this chapter we provide the details of our method of solution for the general object recognition problem. We use the same framework for the automatic recognition of both locations and landmarks within a typical scene, thus the details of our method are presented only for the general object recognition model. One only has to train on images of rooms or locations (see Section 3.5) instead of the set of recognizable objects to use our system for the recognition of locations. While the object recognition problem requires attention to background content (see Section 3.11), in the scene recognition problem, the background content is considered informative and thus unaccounted for.

Section 3.1 reconsiders our problem statement for clarity; Section 3.2 details the mathematical foundations of PCA; Section 3.4 presents the paradigm that we follow for our object recognition model; Sections 3.5 and 3.6 briefly explain the training and testing phases of our method; Section 3.8 details the mathematical foundations of the interest operators we use; Section 3.9 in turn gives details of how chosen interest points are filtered for quality. Section 3.10 shows the processing applied to the local intensity information around the chosen interest points to achieve the robustness desired in our object recognition system and finally Section 3.11 provides the methods we use for the removal of the background content from within the chosen sub-windows.

3.1. Objective

Global PCA-based methods are sensitive to variations in the background behind any objects of interest; the location of the object to be recognized within the image; changes in the orientation of the object; to occlusion or changes in parts of the scene and to changes in scale. 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.

We choose to use PCA (albeit with the disadvantages outlined above) as it is a technique which has proven itself over the years, especially in the face recognition field. From the machine vision perspective it has a very attractive property. It is optimal in a *correlation* sense. If any two images from a set are projected to the eigenspace, the distance between the corresponding points in the eigenspace is a measure of the similarity of the images in the $L2$ norm [**14**]. Section 3.3 provides more insight into the optimality criterion of PCA.

3.2. PCA Overview

PCA allows images to be compared in a lower dimensional sub-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 projected points (eigenpoints) in this space. The smaller the distance between the point representing the image to be recognized and another (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 up to N eigenvectors might be needed to represent the images exactly, a suitably chosen small number of eigenvectors, K ($K \ll N$) is generally sufficient for capturing the primary appearance characteristics of the objects as depicted in typical photographic images [**14**]. Optimal results for such a capture are obtained if the K eigenvectors correspond to the k largest eigenvalues of the covariance matrix $\mathbf{Q}$. Comparisons are thus made in this lower dimensional eigenspace.

The first step in the eigenvector computation procedure requires that the input image set be normalized to ensure that variations in the lighting do not play a role. Given that the input images are normalized, the input set is represented as

$$\{\mathbf{i}_1, \mathbf{i}_2, ..., \mathbf{i}_M\}, \tag{3.1}$$

where M is the number of input images in the training set. Next, an image matrix is created by subtracting the average image $\mathbf{c} = \frac{\sum_j i_j}{M}$ from each input image, and stacking the results in a $N \times M$ matrix

$$\mathbf{P} = [\mathbf{i}_1 - \mathbf{c}, \mathbf{i}_2 - \mathbf{c}, ..., \mathbf{i}_M - \mathbf{c}], \tag{3.2}$$

where M is as before and N is the number of pixels in each image. Note that a constraint is already apparent here as all training images must be of the same size. Note also that subtracting the average $\mathbf{c}$ is useful because it removes the parts of the images that are common to all of them (eg: background if all the training images were taken with the same background), but this step is not absolutely necessary.

To find the eigenvectors that are characteristic of the input images, one needs to diagonalize the $N \times N$ *covariance matrix*:

$$\mathbf{Q} = \mathbf{P}\mathbf{P}^T. \tag{3.3}$$

The eigenvectors $\mathbf{e}_k$ satisfy the equation:

$$\mathbf{Q}\mathbf{e}_k = \lambda_k \mathbf{e}_k, \tag{3.4}$$

where λ_k is the eigenvalue of the k^{th} eigenvector. Note that the $N \times N$ element $\mathbf{Q}$ matrix is huge, since typically there can be 100000 pixels in an image and thus the $\mathbf{Q}$ matrix has a size of 10^{10} elements. Finding the eigenvectors of such a large matrix is impractical.

To account for such a problem various methods exist to compute the eigenvectors. For instance, one can consider the $M \times M$ *implicit covariance matrix* $\tilde{\mathbf{Q}}$ defined as

$$\tilde{\mathbf{Q}} = \mathbf{P}^T\mathbf{P}. \tag{3.5}$$

This matrix is typically much smaller than the $N \times N$ element $\mathbf{Q}$ matrix in practice, as the number of input images M is usually of the order of 100, which is much smaller than $N \approx 100^2$. Given the small size of $\tilde{\mathbf{Q}}$, finding its eigenvectors is a relatively easy procedure, but note that these are not the eigenvectors of the $\mathbf{Q}$ matrix. Fortunately, it can be shown [**13**] using singular values decomposition that the M largest eigenvalues of the $\mathbf{Q}$ matrix and their corresponding eigenvectors are related to the eigenvalues $\tilde{\lambda}_k$ and eigenvectors $\tilde{\mathbf{e}}_k$ of the $\tilde{\mathbf{Q}}$ matrix by:

$$\lambda_k = \tilde{\lambda}_k \tag{3.6}$$

$$\mathbf{e}_k = \tilde{\lambda}_k^{-1/2}\mathbf{P}\tilde{\mathbf{e}}_k. \tag{3.7}$$

Note that the above approach is very useful, since only the first K eigenvectors of the $\mathbf{Q}$ matrix are needed, where $K < M$ and $M \ll N$

As stated above, various methods exist to compute the eigenvectors of $\tilde{\mathbf{Q}}$, specifically, we choose the singular values decomposition [**17**] 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. The singular values decomposition transform serves our needs well as it is numerically very stable and efficient [**17**].

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

$$\mathbf{A} = \mathbf{U}\mathbf{X}\mathbf{V}^T \tag{3.8}$$

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 [**6**]. 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 both the covariance matrix and the implicit covariance matrix are obtained.

The eigenvectors that are found using the above procedure correspond to a basis set that is suitable for the given input images. This can be explained intuitively by considering a simple 2-dimensional example (i.e. $N = 2$) as shown in Figure 3.1, where each point corresponds to an input image. Note that the original basis does not reflect any of the symmetry of the image set, whereas the eigenvector basis is well suited to the image set as the majority of the points in the image lie along one axis of the eigenvector basis. The semi-major axis of the ellipse surrounding the images corresponds to the new eigenvector basis axis with the largest eigenvalue. The description of images with respect to this new coordinate system is much more efficient, as the majority of the points lie along this main axis.

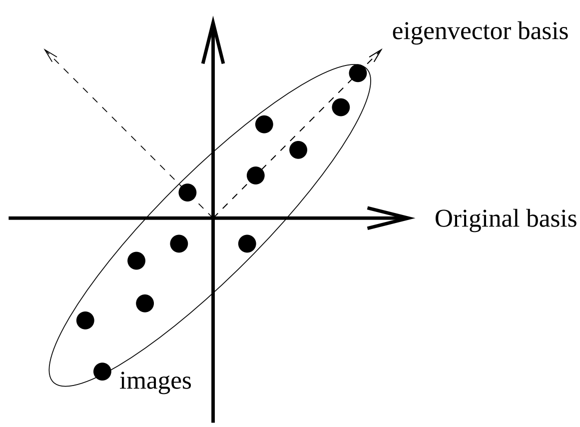


FIGURE 3.1. Example of a 2D eigenvector basis set for a set of input images

Given the eigenvectors, it is now possible to project the input images onto the eigenspace and reduce the storage requirements for the images. Mathematically, the decomposition of an image $\mathbf{i}_j$ onto the eigenvector basis gives a K dimensional vector

$$\mathbf{f}_j = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k]^T(\mathbf{i}_j - \mathbf{c}). \tag{3.9}$$

By projecting all the input images onto the eigenvectors, one gets a set of points in the eigenspace, and the comparison of the images can thus be done in this space.

The eigenspace decomposition of images is very useful from the point of view of storage and computational efficiency. Storage of the approximated images is quite efficient because their eigenspace representation only requires K real numbers, whereas the exact pixel representation requires N real numbers, with $N \gg K$. Now if one needed to compare a new image to the input set, then the new image would first be decomposed onto the eigenspace and then the Euclidean distance to the other images in the eigenspace would be found. Note that the eigenspace representation provides a reasonable approximation of the true distance d between images $\mathbf{i}_r$ and $\mathbf{i}_s$

$$d = |\mathbf{i}_r - \mathbf{i}_s|^2 \approx |\mathbf{f}_r - \mathbf{f}_s|^2. \tag{3.10}$$

For a detailed derivation of the above equation see Section 3.3. The equation implies that the closer the projections are in eigenspace, the higher the correlation between the projected images. Essentially, the advantage of the eigenspace representation depends on two facts:

- Fewer dimensions than the exact pixel representations are required thus efficiency is increased.
- The selected components emphasize *important* variations at the expense of extraneous variations in the images.

3.3. Correlation in Eigenspace

The Euclidean distance between two points in an eigenspace constructed using the K eigenvectors corresponding to the vectors with the K largest eigenvalues, is an optimal measure of correlation between the corresponding brightness images as the distance is an approximation of the *sum-of-squared-difference* (SSD) between the images. We outline that following the work presented in [**14**].

Let $\mathbf{i}_m$ and $\mathbf{i}_n$ be two intensity images with $\mathbf{f}_m$ and $\mathbf{f}_n$ being their corresponding projections in eigenspace. The intensity images can always be expressed as the following linear sum:

$$\mathbf{i}_m = \sum_{i=1}^{N} f_{m_i} \mathbf{e}_i + \mathbf{c} \tag{3.11}$$

where $\mathbf{c}$ is the average of the entire image set; f_{m_i} are the coordinates of the eigenpoint $\mathbf{f}_m$ and $\mathbf{e}_i$ is the i^{th} eigenvector of all the N eigenvectors. Since our eigenspace is constructed with only K eigenvectors corresponding to the vectors with the K largest eigenvalues, our K eigenvectors represent the most *significant* variations within the image set. Thus $\mathbf{i}_m$ can be approximated in the following manner:

$$\mathbf{i}_m \approx \sum_{i=1}^{K} f_{m_i} \mathbf{e}_i + \mathbf{c} \tag{3.12}$$

Immediately we see that the above equation provides the *best* possible linear approximation with fewer than N degrees of freedom as the K eigenvectors capture the most significant variations in the image set.

In machine vision, a technique used extensively for determining the similarity of any two intensity images is template matching. The underlying principle behind such a technique lies in the *sum-of-squared-difference* (SSD) between the brightness values in the images. As $\mathbf{i}_m$ and $\mathbf{i}_n$ are normalized (they are unit vectors), one can relate SSD to correlation in the following manner:

$$\| \mathbf{i}_m - \mathbf{i}_n \|^2 = (\mathbf{i}_m - \mathbf{i}_n)^T (\mathbf{i}_m - \mathbf{i}_n) = 2 - 2\mathbf{i}_m^T \mathbf{i}_n \tag{3.13}$$

where, $\mathbf{i}_m^T \mathbf{i}_n$ is the correlation. We see that maximizing the correlation corresponds to minimizing SSD and thus maximizing the similarity between the images. The left hand side of 3.13 can be re-written in the following manner using equation 3.12:

$$\| \mathbf{i}_m - \mathbf{i}_n \|^2 \approx \left\| \sum_{i=1}^{K} f_{m_i}\mathbf{e}_i - \sum_{i=1}^{K} f_{n_i}\mathbf{e}_i \right\|^2 \tag{3.14}$$

The right hand side of the above equation can further be simplified to give:

$$\begin{aligned} \left\| \sum_{i=1}^{K} f_{m_i}\mathbf{e}_i - \sum_{i=1}^{K} f_{n_i}\mathbf{e}_i \right\|^2 &= \left\| \sum_{i=1}^{K} (f_{m_i} - f_{n_i})\mathbf{e}_i \right\|^2 &(3.15)\\ &= \sum_{i=1}^{K}\sum_{j=1}^{K} \mathbf{e}_i^T\mathbf{e}_j \times (f_{m_i} - f_{n_j})^2 &(3.16)\\ &= \|\mathbf{f}_m - \mathbf{f}_n\|^2 &(3.17) \end{aligned}$$

The last simplification can be made as the eigenvectors are orthonormal; $\mathbf{e}_i^T\mathbf{e}_j = 1$ when $i = j$ and 0 otherwise. Finally we obtain equation 3.10 when combining equations 3.14 and 3.17:

$$d = \| \mathbf{i}_m - \mathbf{i}_n \|^2 \approx \| \mathbf{f}_m - \mathbf{f}_n \|^2 \tag{3.18}$$

Thus the square of the Euclidean distance between the eigenpoints $\mathbf{f}_m$ and $\mathbf{f}_n$ is an approximation of the SSD between their corresponding intensity images $\mathbf{i}_m$ and $\mathbf{j}_n$. The closer the points are in eigenspace, the more highly correlated are the images.

3.4. Recognition Paradigm

Our recognition algorithm consists of a *training* (see Section 3.5) phase and a *testing* (see Section 3.6) phase. During the training phase we run the interest operator (see Section 3.8) on the set of images which we want to recognize (see Figures 4.1 and 4.24). The interest operator returns an interest map consisting of coordinates within the input image that are considered interesting. We proceed to crop square sub-windows of a known size S around a selected group of these interest points (see Section 3.9) and build $\mathbf{P}$.

In the testing phase, we again run the interest operator on the image to be recognized, process the information in the sub-windows (each also of size S) obtained

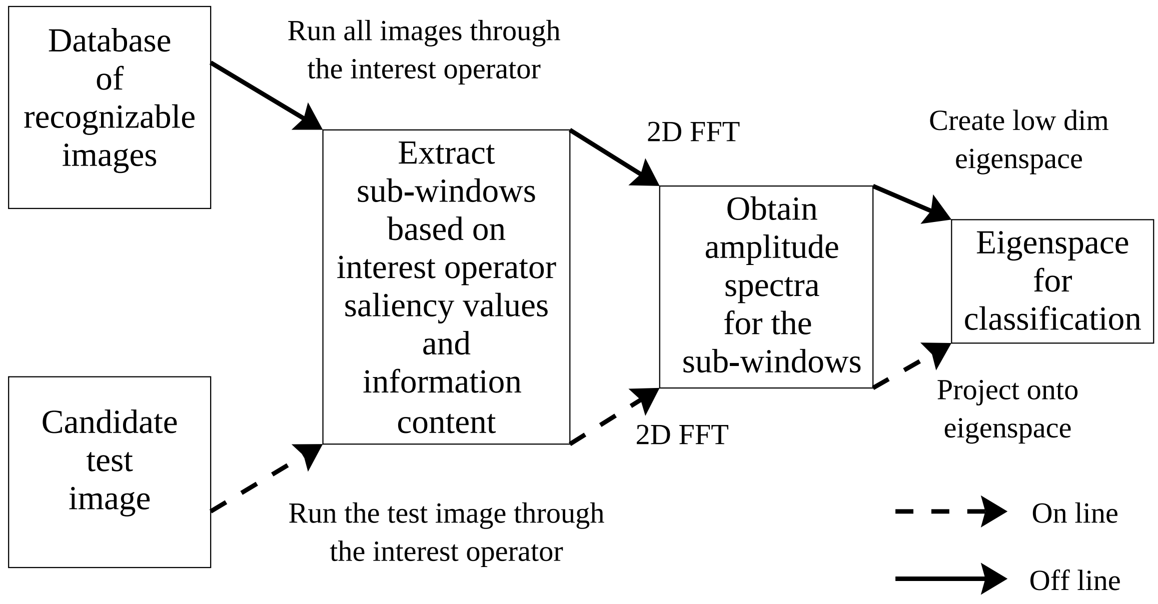


FIGURE 3.2. The recognition model we use

to account for planar rotations and varying backgrounds and project them onto the eigenspace. The classification (see Section 3.7) of the test image is then performed in the already constructed eigenspace. Figure 3.2 shows a flow chart of our recognition model.

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 symmetry 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 3.7), 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.

3.5. Training

The training phase consists of the construction of the eigenspace of recognizable objects. To construct such an eigenspace one first needs to compute the image ensemble **P** as outlined in Section 3.2. In global PCA one constructs **P** using the full set of pixels in every image. Each row of **P** contains the information held by an entire training image. This has numerous disadvantages including the cost of computation, the sensitivity to the background content in the image, and the position of the object within the image. A key aspect of our approach is that we choose areas of interest within the training images and construct **P** using them. This is achieved by first running an interest operator on the set of training images, extracting n sub-windows of a known size S for each training image around the areas considered most interesting by our choice of the interest operator (see Section 4 for the sizes of n and S that we have chosen to use in our experiments). These cropped windows are then used to construct **P**. Hence instead of one image of N pixels being considered as one row of the matrix **P**, n rows of S^2 pixels constitute one training image within **P**[1]. However before these sub-windows are used as the rows of **P**, they are processed (see Section 3.9) to attain the desired robustness of the hybrid PCA and attention operator algorithm.

It should be noted here that when performing general object recognition, the training images consist of the objects to be recognized on a well know background (black in our case, see Figure 4.1). When performing *place recognition* (the term we use for the recognition of rooms or scenes in an environment), the training images consist of recognizable places (scenes) (see Figure 4.24).

3.6. Testing

The on-line testing phase is much simpler. As the eigenspace has already been constructed, upon the arrival of a test candidate image, one simply runs the interest operator on the image, constructs an n by S^2 matrix consisting of the information contained in the n most interesting points returned in the interest map of sub-windows

[1]Note: In general, $nS^2 \ll N$

of size S^2 pixels. This matrix is then projected onto the eigenspace of recognizable images. Following this phase, the classifier proceeds to identify the test candidate image.

3.7. 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. An eigenpoint is defined as a sub-window which has been projected onto the eigenspace. An indexing data structure called 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 to the eigenpoint efficiently, part of our ongoing research involves selecting not just the nearest neighbor but choosing an eigenpoint from the n nearest neighbors.

An approach that seems promising is to choose the appropriate neighbor from the choice of the n nearest neighbors to the projected test sub-window eigenpoint. The choice can be made by examining the n nearest neighbors to that point as a cluster of voters for a sub-image. If the majority of the n nearest neighbors (say there are m of them) are biased against the actual nearest neighbor in terms of the image to which they correspond (i.e. all m of them claim to belong to an image other than the actual nearest neighbor), it is reasonable to assume that the actual nearest neighbor is not the desired match but the nearest of the m neighbors probably is. A concrete example would be the extreme case where the actual (or nearest) nearest neighbor claims to belong to image A while the $n-1$ remaining neighbors claim to belong to image B, the nearest of the $n-1$ neighbors is more likely to be the best match to the projected test image sub-window and the actual nearest neighbor itself, an out-lier. A first order implementation of this was tested, however results were not significantly improved.

Part of our current research involves choosing the correct nearest neighbor appropriately. The results shown in this thesis were all derived by simply choosing the actual nearest neighbor of each eigenpoint.

In order to increase the robustness of the method (due to the erroneous selections of nearest neighbor eigenpoints) a voting mechanism is added. Multiple interest points represent an image or object to be recognized and thus to robustly classify such images one needs to compare all the test image sub-windows to their appropriate training image counterparts. 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 the Euclidean distance.
  - -- Given $\tilde{Y}$ find its corresponding training image T
  - -- Add the value of $1/(D+\epsilon)$ to T's voting 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 4 use an ϵ value of one. Hence a weight of one is added every time a perfect match is found (i.e. one with a value of $D = 0$).

The training image with the largest number of weighted votes W is the designated closest match to the test image. As numerous views of a particular object can be trained for, one can also sum the values of W for images of the same object. For example we use three views per object (when performing object recognition) in our experiments. The best match to a test candidate object is thus the triplet sum of Ws.

3.8. The Interest Operators

3.8.1. Motivation. The complex operations which our brains perform to meaningfully understand our worlds are only inferable on the basis of indirect evidence as a person has no conscious access to the operations themselves. However studies have led to some simple generalizations about visual information processing. Among such generalizations is a two part model scheme where certain aspects of visual processing seem to be accomplished simultaneously (that is, for the entire visual field at once) and automatically (that is, without attention being focused on any part of the visual field). Other aspects of visual processing seem to depend on focused attention and are done serially (one at a time), as if a mental spotlight were being moved within the visual field [**28**]. David C. Marr suggested a first stage of processing (called the *primal sketch*) whereby the pattern of light reaching the visual field is incident on an array of receptors which convert the information into a coded description of lines, spots or edges and their locations, orientations and colors. The actual representations of surfaces and volumes and the final identification of the objects themselves could begin only after this initial coding.

Using this biologically motivated two part model where in the first stage, features are extracted from the patterns of light followed by the second stage where the identification itself is performed, we implemented our object recognition scheme. We extract the features using interest operators, followed by the actual recognition of the object itself using the extracted features. We implemented a means to do this where the attention operator itself is considered independent of the object recognizer, thus allowing us to consider any attention operator as a simple, independent processing *module*. We further added the ability to search for specific attention points based on the *signature* of the attention point. We define such a signature as an *n-tuple* of information returned by the interest operator for the saliency measure(s) of an interest point. For example, an edge density based operator could return points with saliency values proportional to the edge density at those points together with information on say the error margin of that saliency value, thus its signature could simply be the *2-tuple* of edge density and associated error. Such an ability allows for the search of interest points in test images, based on the signatures of the interest points chosen in the training phase thus allowing a level of discrimination when choosing interest points from an interest map returned by the operator.

We considered three different attention operators for our experiments. An edge density and edge orientation [**3, 4**] based operator was used for our first order tests. While the edge density information provided reasonable results, the edge orientation information did not. The operator itself is fast and it provided useful insight into the appropriate selection of features for recognition. The most successful, while being the slowest, was the symmetry based attention operator developed by Ariel Tankus[2] at Tel-Aviv university using a symmetry operator developed by Reisfeld, Wolfson and Yeshurun [**18**]. Finally we also tried the KLT (Kanade, Lucas and Tomasi) [**26, 22**] operator which provided the second best results.

3.8.2. Density and Orientation Operator. In [**3, 4**] Bourque and Dudek present an attention operator based on a notion that attention is drawn to regions that are statistically anomalous (motivated by extensive psychophysics to this effect). They look for regions that differ from the typical edge element distribution within single images based on the assumption that humans tend to concentrate on regions that are unusual or distinctive. They begin with an edge detection process on the image of interest. This returns two maps: an edge map containing edge elements and their associated intensities, as well as an orientation map containing the orientations corresponding to the edge elements. The authors then identify attributes within images using these two maps that allow them to rapidly identify regions of interest and among these are the edge element density and the edge orientation. The edge element density, they define as the extent that the edge density in a local neighborhood

[2] http://www.math.tau.ac.il/~arielt/

differs from the mean density of the entire image. The edge orientation, they define as the extent to which the edge orientation in a local neighborhood differs from an estimate they compute of the most likely orientation within an image.

Maps are, in turn, created for both the edge element density and orientation as defined above. The extrema of these maps, that is the locations with the highest deviations from the mean for the density case and the highest deviation from the maximum orientation for the orientation case, are then found and are considered to be the areas of interest. The authors also provide a means to combine the information from the two criteria discussed above for the selection of interest points.

We used the edge density and orientation operator implemented by Bourque for the tests of our object recognition system. While the results were not as good as those returned by the symmetry operator, they provided insight into the fact that attention points returned near the periphery of the objects under scrutiny tended to be more unstable due to the inclusion of background information in the sub-windows cropped around such points. This in turn led to the development of the masking approach to account for the background information within the sub-windows, presented in Section 3.11.

3.8.3. Symmetry Operator. In [**5**] an excellent overview of attention mechanisms in computer vision written by Yeshurun can be found. He states that attention is without a doubt a concurrent top down *and* bottom up process. The detection of regions of interest are both context dependent (top down) as an attention system can be task oriented, and context free (bottom up) as attention systems should be able to detect and handle unexpected signals. The article proceeds to outline various attention mechanisms in computer vision stating that the first step in practically every artificial vision system begins with some form of edge detection. The symmetry operator [**18**] that we used similarly starts with such an edge map.

Yeshurun *et al.* introduce symmetry based attention mechanisms with the motivation that symmetry attempts to encompasses a broad concept that generalizes most of the existing context free methods for the detection of interest points. Natural and artificial objects often give rise to the human sense of symmetry. While this sense of symmetry is more general than any mathematical formulation (for example a human face is considered highly symmetric by your average human observer although there is no *strict* reflection-al symmetry between both sides of a face), one can focus attention in machine vision based on such a notion of symmetry. The symmetry transform [**18**] we use is such an operator as it attempts to algorithmically formalize the intuitive notion of symmetric image content. It assigns a *symmetry magnitude* and a *symmetry orientation* to every pixel in an image at a low level vision stage following an edge detection process. They compute a *symmetry map*, which they describe as a new kind of edge map, where the magnitude and orientation of an edge depends on the symmetry associated with the pixel. Based on the magnitude and orientation of the symmetry map, they compute the *isotropic* symmetry and the *radial* symmetry (which emphasizes closed contours). They consider strong symmetry edges to be natural interest points, while linked lines are considered as symmetry axes.

The bulk of our results presented in Chapter 4 were computed using the context free symmetry operator due to its success from a recognition perspective. It is invariant under planar rotations and translations, thus providing an excellent selection of interest points within an image, irrespective of the planar orientation of the object within the image. As our framework for recognition is also formulated to be planar rotation invariant, we assume that the attention operator returns similar points for both the training case (where the object is upright) and the test case (where the object can be rotated planarly). This operator has this desired property. Furthermore the points chosen usually lie well within the object (as opposed to the periphery) thus bypassing the need to mask over the background content which can be found in sub-windows cropped around the borders of the objects (see Section 3.8.2).

3.8.4. KLT Operator. In [**26, 22**] an interest operator is proposed based on the notion of detecting and tracking robust features in a sequence of images. While the operator is primarily designed for tracking features, it provides an excellent means of detecting important corners in an image. The authors define the quality of their features by stating that good features are those which can be tracked well thus their selection criteria is optimal for their needs. In practice they compute the eigenvalues of the following 2x2 matrix G :

$$G = \sum_{W} gg^{T} \tag{3.19}$$

where g is the image gradient $(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})$, I is the image intensity and W is the window over which it is computed. G has two eigenvalues λ_1 and λ_2. Two small eigenvalues mean a roughly constant intensity profile within a window. A large and a small eigenvalue correspond to a unidirectional pattern and two large eigenvalues represent corners, salt-and-pepper textures or any other pattern that can be tracked reliably [**26**]. In practice the window W is selected if the minimum of the two eigenvalues is larger than that of a predetermined threshold λ.

In practice this operator despite being motivated for tracking provided good results in our experiments and was also very attractive due to its speed in returning interest points.

3.9. Filtering Interest Points

Once an interest map is obtained from the attention operator, interest points are chosen from the map such that the information 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.

The image content around each selected interest point is extracted in a manner that is not purely local. We wish to place an emphasis on the immediate neighborhood of the interest point chosen, while image data as one heads to the periphery of the sub-window, we wish to consider less important. We achieve this by multiplying the data within the sub-window with a two dimensional Gaussian centered at the chosen interest point. This reduces sensitivity to distal points which may be on the background while emphasizing the immediate neighborhood of the selected interest point. This also reduces the sensitivity to the shape of the window. Advantages are also gained for the achievement of rotation invariance as outlined in Section 3.10.

3.10. Feature Processing

Once sub-windows are cropped around the chosen interest points, the intensity information within these local areas of interest are processed such that they are invariant to planar rotations. The use of a Fourier basis for the sub-windows chosen around the interest points provides such a rotation invariance.

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:

$$e^{j2\pi(au+bv)}F(u, v) \tag{3.20}$$

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:

$$|e^{j2\pi(au+bv)}F(u, v)| = |F(u, v)| \tag{3.21}$$

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

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

FIGURE 3.3. 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

shifts in θ in the polar image, see Figure 3.3. 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. Some of the high frequency components in the spectrum introduced due to the discontinuities in the discrete two dimensional image which we treat as a continuous signal are thus suppressed. Section 3.9 also outlines advantages of such an operation.

We then proceed to obtain the amplitude image using 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 of the full image about arbitrary points are handled automatically since for any interest point they can be described as a translation and a rotation about the center. The phase image returned by the Fourier transform is currently discarded.

3.11. Masking

Interest points chosen by an operator can sometimes lie on the periphery of the desired object. Interest operators which rely heavily on the boundaries of objects for example, can return interest maps where the more salient points occur precisely on the boundaries of the objects we are trying to recognize. When such a sub-window is projected onto the eigenspace, as a large amount of the information in the sub-window can contains an unknown background, the projected point will not necessarily lie near (according to an $L2$ norm) any of the projected training image sub-windows.

We have developed a technique to handle such cases, however it is currently restricted to cases where the test image being recognized exhibits no rotations of any sort (planar rotations are not accounted for).

Given that all our training image samples (for the object recognition problem) contain recognizable objects on a known black background, the idea is to directly overlay training image sub-windows onto a given test image sub-window and to *mask* over the potential background pixels in the test image sub-window prior to projection onto the eigenspace. The masking is simply done by coloring pixels in the test image black in accordance to their corresponding training image pixels. So given a test image sub-window, if the closest corresponding training sub-window image in the eigenspace is greater than a predefined threshold for what is considered a good match in the eigenspace, then for each training sub-window, overlay that window onto the aberrant test sub-window, apply the mask determined by the training window onto the test window and re-project the masked test window onto the eigenspace. Compute the distance to the nearest neighbor in the eigenspace and store that value. The idea is to choose the masked projection which provides the shortest nearest neighbor distance in the eigenspace as this should correspond to the training sub-window which best matches the test image sub-window irrespective of the background on the test image. Clearly this technique only applies to points chosen on the periphery of an object to be recognized and that the test image sub-window has not been processed using the techniques proposed in Section 3.10. Results show significant improvements when using such a mask, however the task is expensive due to the number of projections onto the eigenspace (one for every training image sub-window) and is restricted to test images where it is known *a priori* that the object to be recognized has not rotated substantially.

CHAPTER 4

Experimentation and Results

This chapter presents our experimental results for two types of recognition tasks, object recognition and place recognition. Our object recognition experiments are presented in section 4.1 and our place recognition experiments appear in section 4.2. At the end of section 4.1, we present an independent set of experiments to highlight the usefulness of our masking technique detailed in section 3.11.

Figure 4.1 contains the set of objects used for our object recognition experiments, Figure 4.13 contains a small set of recognizable objects used for our masking experiments (these are taken from the COIL image database at Columbia university) and finally Figure 4.24 shows the set of rooms for training our place recognition experiments. All the images used for training and testing (bar those for the masking experiments) were created in our lab[1] due to the unavailability of public image databases suitable for our experiments[2]. Image databases such as Columbia University's COIL database, despite having an excellent variety of objects suitable for recognition, contained restricted backgrounds and were thus unsuitable for our purposes. We however did use these for our preliminary object recognition experiments with great success and furthermore to demonstrate the usefulness of masking (Figure 4.13). All the images used for object recognition are large (640x480) color images while the ones

[1]These images are available to other researchers at http://www.cim.mcgill.edu/~dudek/objects
[2]In particular, we needed a database with both rotated objects and background variation

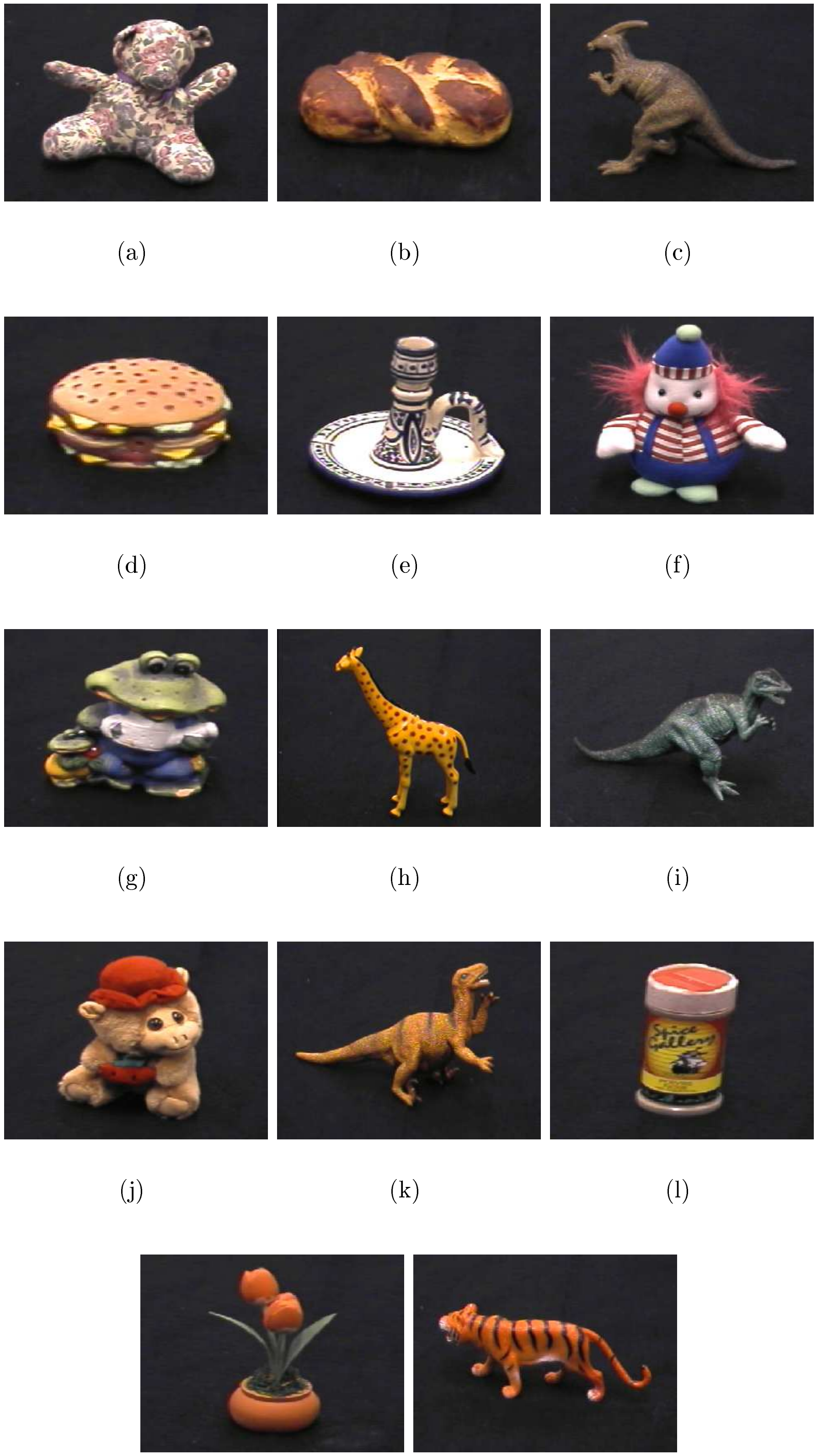
used for place recognition are smaller (256x256). Note that the images shown in this thesis have been cropped and scaled appropriately for legibility.

For the object recognition experiments, while care was taken to keep the notion of scale between test and training images constant, the experimental setup was rather crude. A simple color digital camera was chosen to take pictures of the set of training objects on a matt black background. The same objects were then placed on naturally occurring backgrounds in our lab at *more or less* the same distance from the camera, thus obtaining the test image samples (see Figure 4.5). Due to the robustness of the system created, our results show that the experimental environment need not be specifically tailored to provide good results. While we were concerned with lighting, no special effort was made to construct an environment with constant lighting and despite that our results are good.

For our place recognition experiments, we took pictures of various rooms on our floor using a hand held digital camera. Test images (Figures 4.25,4.26,4.27,4.28) were similarly taken at slightly different locations in the same rooms with the camera aligned at different orientations. Note again that there was no tripod involved and that despite the crudeness of the experimental setup the results were very good.

4.1. Object recognition

We first present the improvements offered by our approach of performing PCA for object recognition using only local sub-window information. Figure 4.2 shows the test image which we will use to show such an improvement. To first show the advantage of using local sub-window information over the classic global PCA techniques we consider the results obtained by first cropping around the object of interest (see Figure 4.3(a)). The corresponding best match is also shown in Figure 4.3(b). Sub-windows immediately show an improvement as the best match returned when one uses them, see Figure 4.4(a), is the correct one. Note that despite choosing a window appropriately for the case shown in Figure 4.3, there is enough background present within the image to falsely classify the input test image despite the area around



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n)

FIGURE 4.1. Database of training objects (all 640x480 images)

(a)

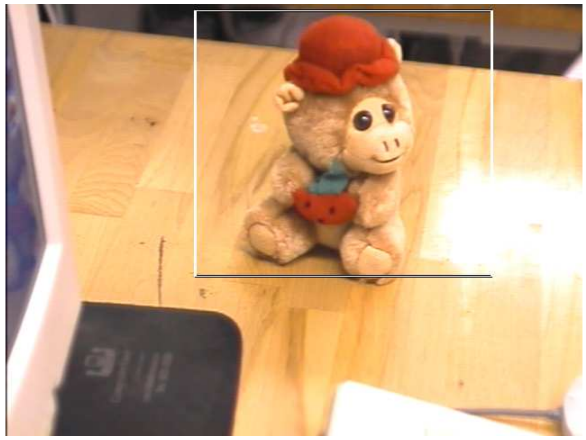
FIGURE 4.2. Test image used for demonstrating the advantages of a sub-window scheme

the center of that window being more emphasized due to the Gaussian weights (see Section 3.9).

The bulk of our object recognition experiments were done using the symmetry based attention operator (see Section 3.8.3). As the points are chosen on a basis of local symmetry, the points returned in test images are very similar to those obtained in the training phase, irrespective of translations and rotations of the object under scrutiny within the test image.

We evaluate the recognition performance of our system on the training set presented in Figure 4.1 using sub-window sizes of 5x5, 10x10, 15x15 and 20x20 pixels respectively. The eigenspace itself was created using 64 interest points per training image and the 30 most significant eigenvectors were used for classification. Tests were conducted to evaluate the accuracy of recognition as a function of the number of interest points chosen in the testing phase.

All the images shown in Figure 4.5 were used as test samples to obtain the results shown in Figure 4.6. Note that these test images include training objects 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



(a)

(b)

FIGURE 4.3. Test image case with one window encompassing the object of interest, followed by the best returned match

was still reliable. Note also that some of the objects were partially occluded from the camera view. While these test images were used for the results presented in Figure 4.6, various other test images were used for the evaluation and performance of rotation invariance recognition and to contrast how different attention operators provide different ranges in the quality of their results.

Figure 4.6 shows two performance plots. The first graph shows the percentage of images that were recognized plotted against the number of interest points used as

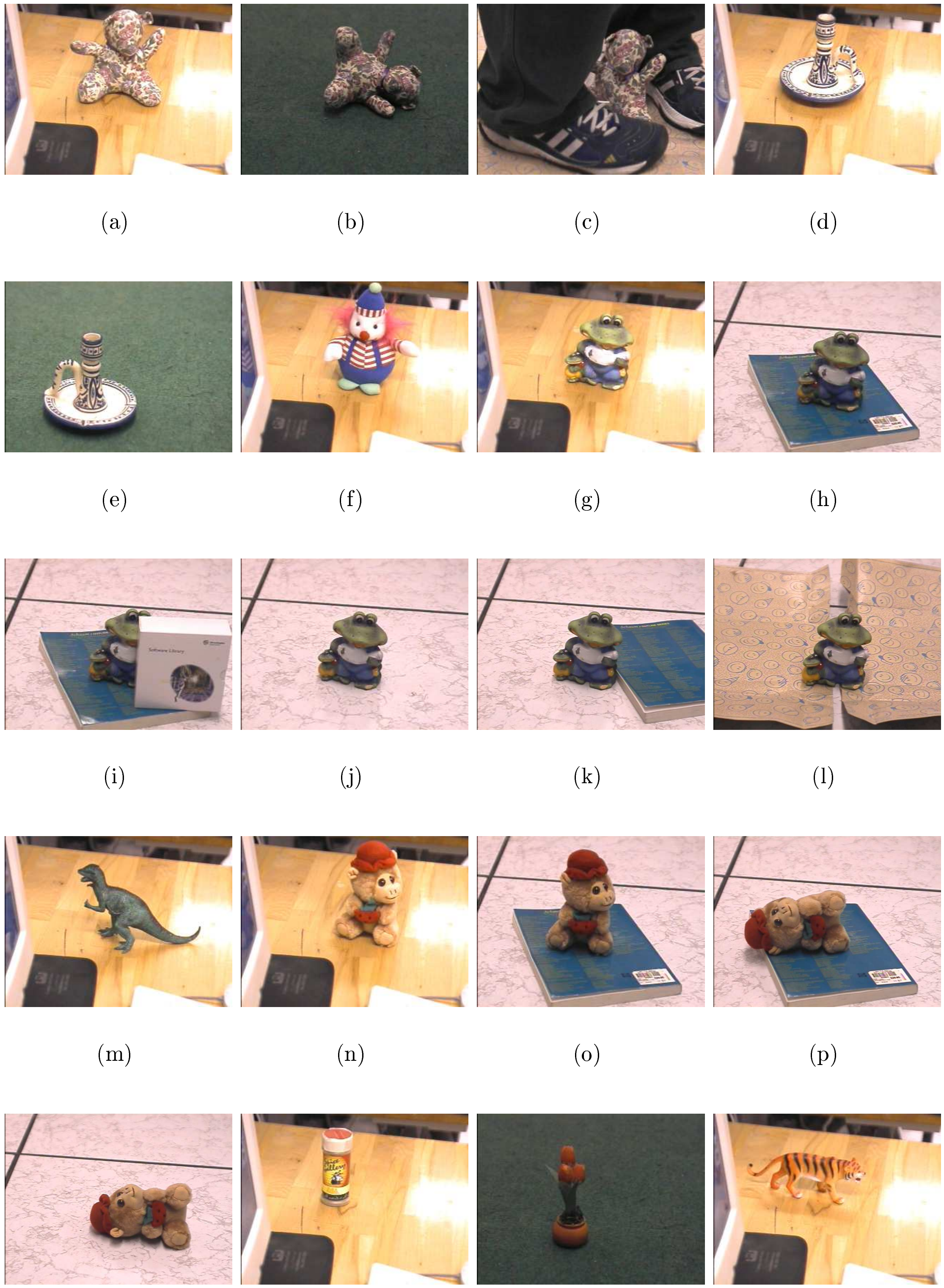
(a)

(b)

FIGURE 4.4. Test image case with 50 windows chosen on the object of interest (using a symmetry based operator), followed by the best returned match

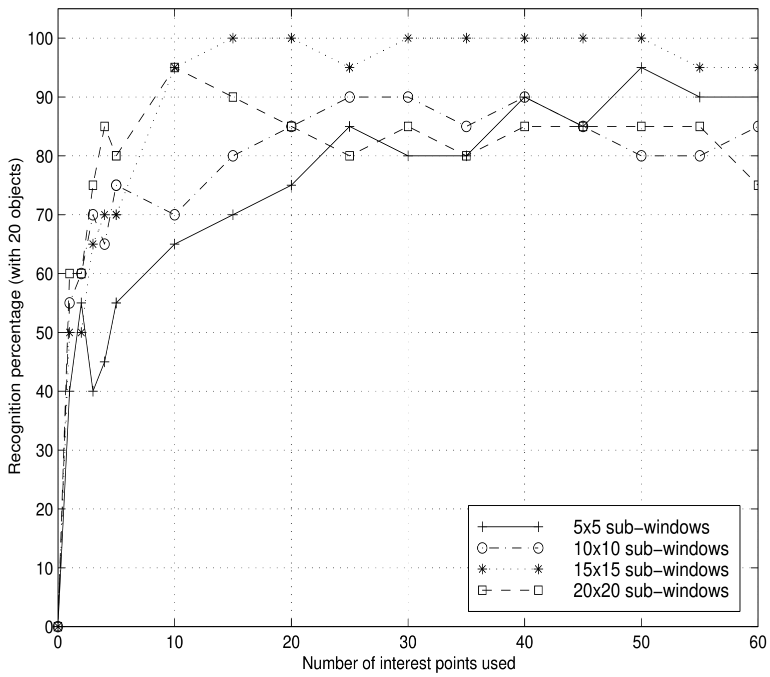
input. Note that with 15 interest points all of the objects shown in Figure 4.6 are recognized with a sub-window size of 15x15. The second graph shows an increase in the reliability of the recognition as one increases the number of interest points.

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 3.7) between the recognized object and the second best

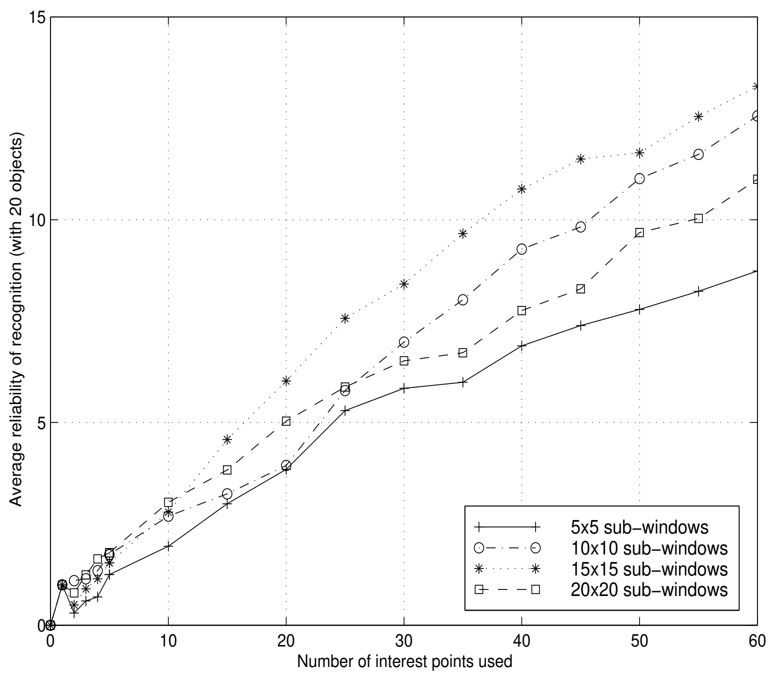


(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o) (p) (q) (r) (s) (t)

FIGURE 4.5. Test images (all 640x480 images) used for the results shown in Figure 4.6.



(a) Recognition rate as a percent vs number of interest points



(b) Reliability of recognition rate vs number of interest points

FIGURE 4.6. Results obtained when varying the number of interest points on images shown in Figure 4.5. The first graph shows the recognition performance on all the 20 images in Figure 4.5 as a percentage. The second graph shows the reliability of the recognition performed. See page 39 for a definition of reliability.

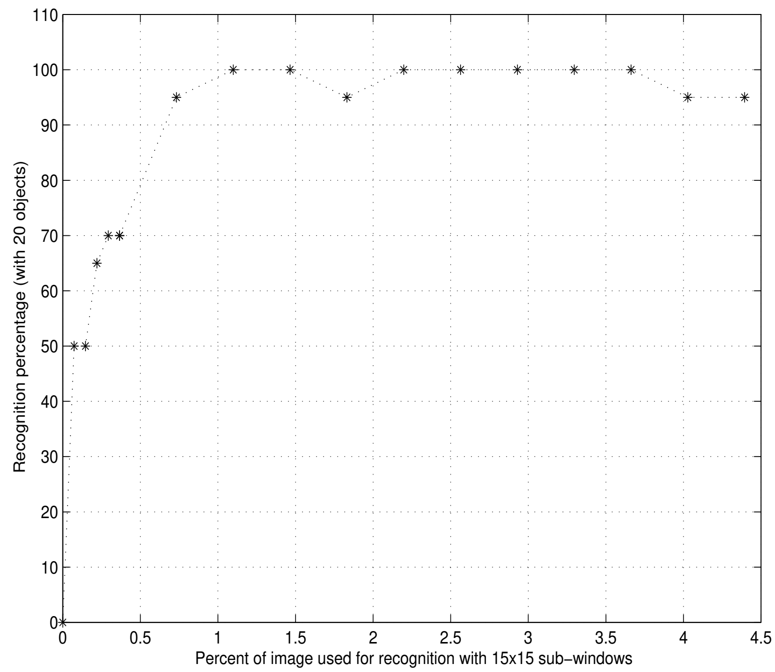


FIGURE 4.7. Results obtained when varying the number of interest points on images shown in Figure 4.5. The x-axis shows the percent of the image used instead of the number of interest points for the 15x15 window size.

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 0.16% of a 640 by 480 pixel image (see Figure 4.8). 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. Figure 4.7 shows the results obtained for recognizing the images using sub-windows of size 15x15 but instead of showing the number of interest points used on the x-axis, we show the percentage of the image used. Note that we can reliably recognize all the images using less than 5 percent of the image, a fact which seems surprising.

It is interesting to note that the recognition percentage at times drops due to an increase in the number of interest points chosen (see Figures 4.6, 4.7). This can be explained when one considers the fact that as one requests more interest points, the

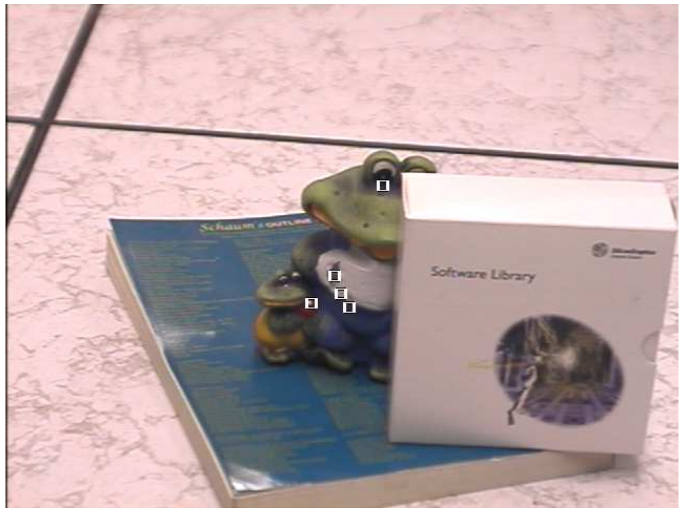
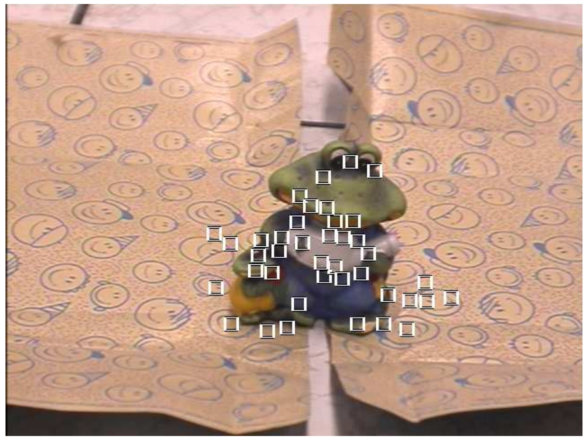


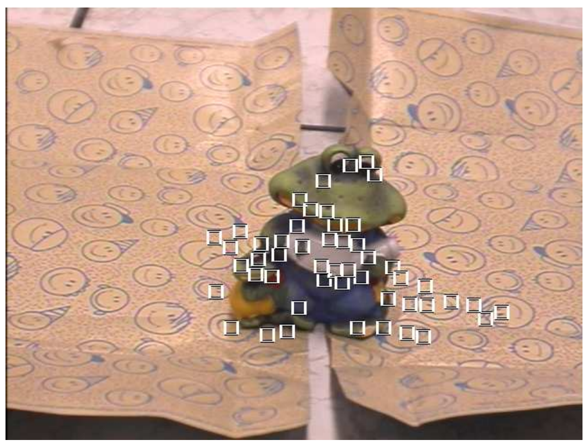
FIGURE 4.8. Example of a test image which was recognized with only 5 interest points with sub-window sizes of 10x10, when using the symmetry operator.

probability of choosing interest points that are outliers, i.e. lie on the background instead of on the object, increases. This is due to the larger amount of background content in the image as the object occupies less than half of the image. Figure 4.9 exemplifies this. It depicts the case of the image which is misclassified when using 50 sub-windows but correctly classified when only using 40 (for the 15x15 sub-window size case). Note that there are more sub-windows which fall on the background when 50 are chosen, and these outliers cast more false votes, mis-classifying the object. We have considered various methods to account for these outliers and they are detailed in Section 5.2.

While the training phase is time consuming due to the construction of the eigenspace, the actual projection of the extracted sub-windows onto the eigenspace followed by their classifications can be performed fast. For the 15x15 sub-window results presented in Figure 4.6, approximately 15 hours were required for the computation of the 30-dimensional eigenspace when constructed using 64 sub-windows for each of the 42 images (three views for each object shown in Figure 4.1). The projection and classification of any given test image once 64 sub-windows had been extracted was



(a) 40 sub-windows



(b) 50 sub-windows

FIGURE 4.9. Test image case with 40 sub-windows chosen on the object of interest (using a symmetry based operator), followed by one with 50 sub-windows. Note that there are more outliers when 50 sub-windows are chosen

performed in approximately 2 seconds. These metrics were obtained on a Pentium II running at 400 MHz with 256 Megabytes of RAM. While our code was written efficiently in C++, it was not optimized for speed, thus performance could radically be improved with some minor optimizations. We did not include the time taken for the extraction of the interest points in the metrics presented above as this varies from operator to operator and is also dependent on various operator specific factors, for example, the scale at which one chooses to run the operator or the size of the image on which the operator is being run.

An analysis of how the chosen window sizes affects recognition is presented here while still part of our ongoing research. From the results shown in Figure 4.6 we clearly see that sub-windows of size 15x15 provide the best results as all the objects in Figure 4.5 are correctly identified with 15 interest points. However we also see that these results are not as *reliable* (as defined earlier) as those obtained when we use 50 interest points. While it is clear that increasing the number of interest points increases the reliability of the recognition, the percent of recognized images plateaus quickly when one uses around 30 interest points. The reliability plots show that an increase in the window size also provides an increase in the reliability of the recognition but immediately we see that while an increase in the window size provides more information for the purposes of recognition, this is not necessarily a good choice to make. Using larger and larger sub-windows is clearly not the answer however due to the fact that the larger the sub-window becomes, the more likely it is to contain background information, simply due to the following scale problem. One needs *a priori* knowledge of the size of the object to be recognized within an image to suitably choose a window size. A sub-window size must clearly be chosen specific to the size of the object which we wish to recognize if it is to encompass a minimal amount of the background, especially if edge based attention operators are to be used. However this process is legal as the sub-window sizes are chosen during the training phase where human intervention is allowed and thus sub-window sizes can be appropriately tuned. In a model where scale invariant features are extracted, this should pose less of a problem and this is also part of our current ongoing research. Figure 4.10 shows an example where the use of a large sub-window (20x20) results in a reliably recognized object, while smaller sub-windows fail to do so.

The object shown in Figure 4.11 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. Note that this is specific

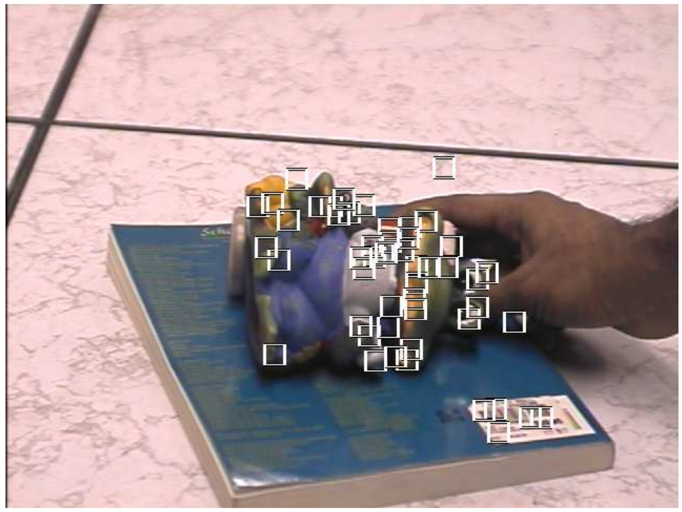


FIGURE 4.10. Example of a test image which was reliably recognized using 50 interest points and a window size of 20x20 but failed for the 10x10 case when using the symmetry operator

to the symmetry operator as the points chosen on the background are in response to the salient symmetric features present there. Another 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 (the symmetry operator) as the scale over which it was working was not appropriately tuned. The image shows 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.

4.1.1. The interest operators. While the bulk of our results presented were obtained using the symmetry operator, we conducted experiments with two other interest operators. While the symmetry operator provided the best results, it was by far the slowest. The other operators had their own advantages in terms of the interest points chosen and the speed at which they chose them. As the images used for object recognition were all large 640x480 images, fast interest operators were beneficial for a first order evaluation of our recognition system due to temporal

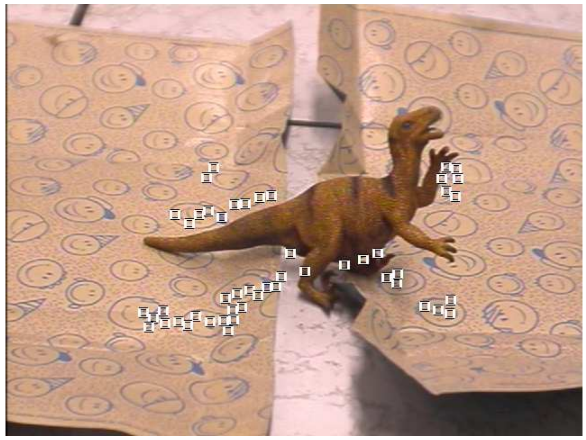
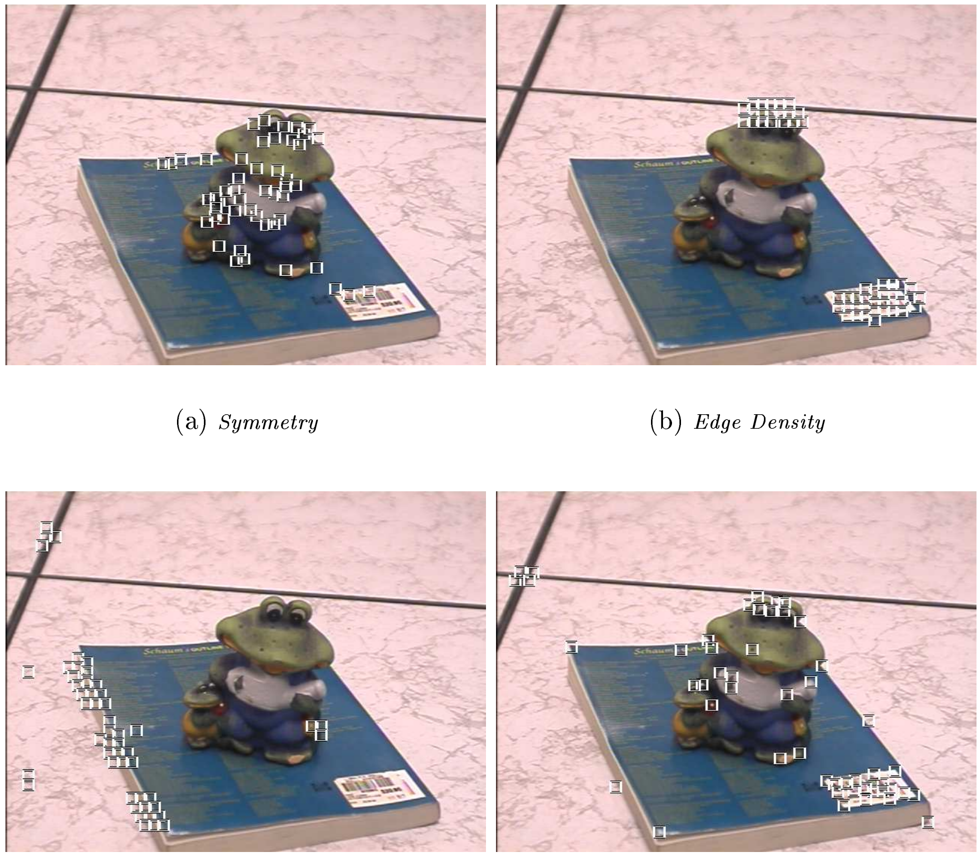


FIGURE 4.11. Example of a test image which was not recognized using 50 interest points when using the symmetry operator. Note that the majority of the interest points chosen fall on the highly symmetric background

constraints. The operators we considered in the order of performance from our object recognition perspective were the symmetry operator [**18**], the KLT operator [**26, 22**], and an edge density and orientation based operator [**3, 4**]. The edge density and orientation operator provided comparable results to those of the KLT operator when one considered only the edge density of the images. The KLT operator was the fastest but due to the criteria of selection used in [**26, 22**] which revolved around a tracking paradigm, the points chosen were not as consistent between training and test images of the same object. Figure 4.12 depicts the behaviors of the interest operators we used. It should be noted that as the operators had a variable number of parameters which could be tuned for improved performances, a thorough comparison of their performance is beyond the scope of this thesis. Figure 4.12 is merely a depiction of their performances on an exemplary test image. Given the parameters we used for our experiments, the symmetry operator provided the best set of interest points as the majority of them were chosen within the object of interest.

It is interesting to note that when interest points are returned within the object of interest, the background becomes irrelevant. Successful recognition can be performed without considering the outlines of the objects being recognized and without



(a) *Symmetry* (b) *Edge Density*

(c) *Edge Orientation* (d) *KLT*

FIGURE 4.12. Here we show the results of choosing 50 interest points using the four operators we considered on one of our test images.

any segmentation of the object. While the importance of an object's outline for recognition is indisputable, especially in the stages of early visual processing [**2, 12**], our experiments demonstrate the ability to recognize objects merely based on the colors and textures of the objects.

4.1.2. Masking Experiments. We now present an independent set of experiments used to explicitly demonstrate the advantages of our masking technique (see Section 3.11). The images in Figure 4.13 were used to construct an eigenspace of recognizable objects. We then placed the recognizable objects on non-uniform backgrounds creating the images shown in Figure 4.14 such that we could test the masking

(a) *oralb bottle*

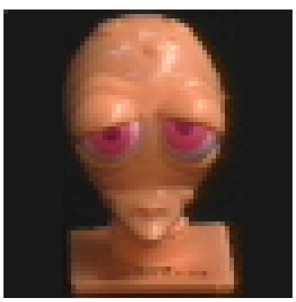
(b) *porcelain cat*

(c) *wooden mushroom*

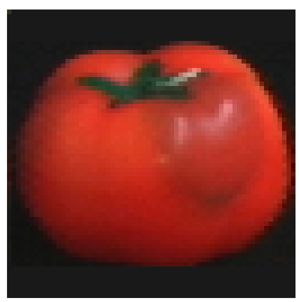
(d) *plastic bottle*

(e) *plastic cat*

(f) *rolaid bottle*



(g) *alien head*



(h) *tomato*

(i) *teacup*

FIGURE 4.13. Training images used for our masking experiments (taken from the COIL image database)

algorithm. The images in Figure 4.14 were used as a test images and the task was to correctly locate and identify the objects within the test images.

Our first task was to locate the recognizable objects within the test images. Turk and Pentland [**30**] use PCA itself to find faces within an image for the purposes of

(a) *Green background*

(b) *Paper background*

(c) *Slate background*

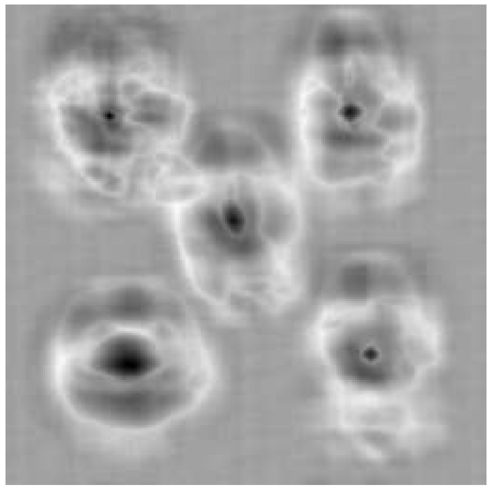
(d) *Complex background*

FIGURE 4.14. Images used as test images. Note the non-uniform backgrounds.

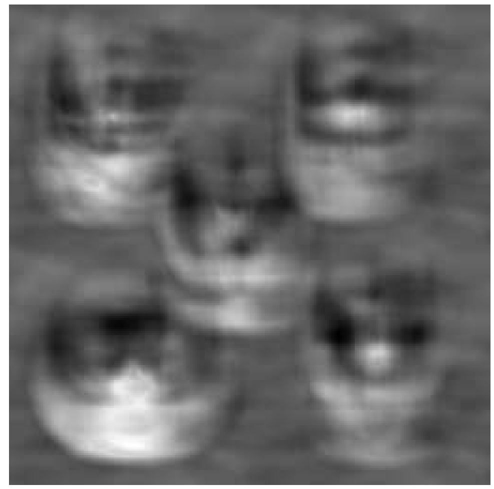
face recognition. Similarly, we use PCA here to find the objects of interest within a test image. Given a test image, we find areas of interest within the test image by performing PCA at various locations within the test image. One can think of this as using a PCA based interest operator to locate interesting areas within a test image. Specifically this is done by sampling training image sized windows (64×64) from the test image and projecting those *sub-images* onto the eigenspace constructed using the training image set. Once projected, the smallest value of d (ie. the distance to the closest training image in the eigenspace) is stored for the pixel corresponding to the center of the *sub-image*. Note that this can be done for every pixel in the training image if required. Areas of interest then correspond to pixel locations containing the smaller values of d. Figure 4.15 shows our results when performing this task. The values of d found for the test images shown in Figure 4.14 are presented in the images where the smallest values of d are mapped to black and the largest to white, thus the darker areas indicate areas of interest.

Once areas of interest are found, we proceed to crop around the area. For each of the cropped images we first superimpose the training images (one at a time) and color in the tentative background pixels according to the training image templates. Once the mask from a training image is applied, the resultant image is projected onto the eigenspace and the closest training image eigenpoint is found. If the distance d to that eigenpoint decreases for the pixel (corresponding to the center of the sub-image), we store the new lower value for that pixel and the closest corresponding training image eigenpoint. Once this is performed for all the pixels in the neighborhood of the areas of interest, the pixel with the lowest value of d represents the most likely place where an object to be recognized is residing and the corresponding eigenpoint (which provides the value of d) is that of the object residing at that location.

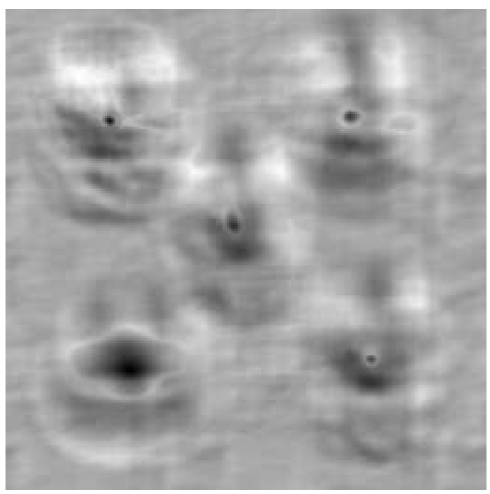
Figures 4.17 to 4.21 present results obtained when using Figure 4.14(a) as the test image. The points of interest (stated in the top rows of all the tables) are obtained without any background extraction and are the five most tentative areas of interest (sorted by increasing values of d). Every pixel within a 64×64 *sub-window* centered at that point of interest is masked and re-projected onto the eigenspace. Each row corresponds to the training image that was used to extract the background (the masking template), and the entry for d corresponds to the best (lowest) value that was obtained in the entire *sub-window* when the background was extracted using that training image. Note also that the corresponding (x, y) coordinates of where these matches are found are also listed. Figure 4.16 contains the coordinates corresponding



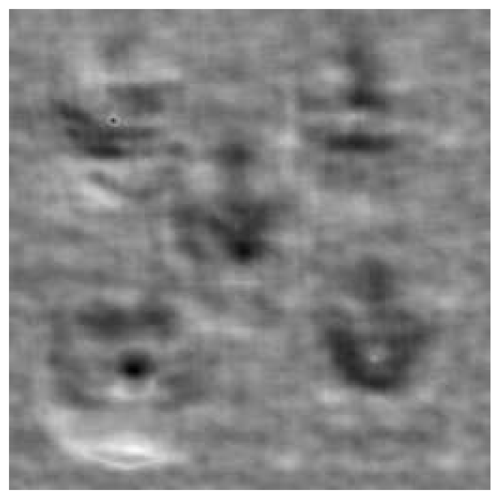
(a) *Green background*



(b) *Paper background*



(c) *Slate background*



(d) *Complex background*

FIGURE 4.15. Eigenspace distance maps corresponding to images in Figure 4.14. The smallest d values are mapped to 0 (black) and the largest to 255 (white).

to the centers of the training images once they are placed on textured backgrounds and is useful to the reader in interpreting the validity of the results obtained.

Finally Figure 4.22 shows the accumulated results for all the test images shown in Figure 4.14. Each row of Figure 4.22 contains the minimum value of d at each location

Object to recognize	x coordinate	y coordinate
Porcelain cat	55	59
rolaid bottle	183	57
alien head	121	115
tomato	66	192
plastic cat	194	186

FIGURE 4.16. Table of object coordinates. Note that these are identical for all the training images (compare with Figure 4.14) regardless of the backgrounds. The coordinates pinpoint the centers of the various objects.

Examination of neighborhood of pixel $(66, 192), d = 0.0160325$ without masking			
x coordinate	y coordinate	minimum eigenspace distance (d)	masking image
64	161	0.0336721	oralb bottle
64	161	0.0482151	porcelain cat
66	192	0.0319262	wooden mushroom
64	161	0.0480483	plastic bottle
79	200	0.0616677	plastic cat
36	169	0.0660311	rolaid bottle
64	160	0.0400758	alien head
66	192	0.000329012	tomato
66	192	0.0118267	teacup

FIGURE 4.17. This table shows the minimum distances found in the 64×64 neighborhood of pixel $(66, 192)$ for each applied mask. Note that the *tomato* is clearly the object at that location as the tomato mask provides the smallest value of d.

and the columns depict the masks which provided those values. Masking clearly improves the classification of objects on non uniform backgrounds in the constructed eigenspace as the distances obtained to the nearest neighbors generally are an order of magnitude less than those obtained without masking.

4.1.3. Color. Finally we would like to highlight the importance of color in our experiments. Remarkably we see that the results obtained when using only gray-scale images (see Figure 4.23) are poor. Figure 4.23 was obtained using the same parameters as those used for Figure 4.6. We show the plot obtained when using sub-window sizes of 15x15 pixels as these provided the best results when color was used. Only 10 of the 20 objects are recognized at best (when using 50 interest points) and the reliability of even that success rate is low.

Examination of neighborhood of pixel $(55, 59), d = 0.0222514$ without masking			
x coordinate	y coordinate	minimum eigenspace distance (d)	masking image
55	59	0.0052982	oralb bottle
55	59	0.00121013	porcelain cat
55	59	0.0303485	wooden mushroom
55	59	0.0116977	plastic bottle
55	59	0.0545919	plastic cat
55	59	0.0422481	rolaid bottle
55	59	0.0314643	alien head
55	59	0.056568	tomato
74	60	0.0378017	teacup

FIGURE 4.18. This table shows the minimum distances found in the 64×64 neighborhood of pixel $(55, 59)$ for each applied mask. Note that the *porcelain cat* is clearly the object at that location as the porcelain cat mask provides the smallest value of d.

Examination of neighborhood of pixel $(121, 115), d = 0.0322604$ without masking			
x coordinate	y coordinate	minimum eigenspace distance (d)	masking image
121	115	0.0102841	oralb bottle
121	115	0.017435	porcelain cat
121	115	0.0295266	wooden mushroom
121	115	0.00857163	plastic bottle
121	115	0.0106105	plastic cat
121	115	0.0101183	rolaid bottle
121	115	0.00209664	alien head
121	115	0.078991	tomato
138	119	0.0500167	teacup

FIGURE 4.19. This table shows the minimum distances found in the 64×64 neighborhood of pixel $(121, 115)$ for each applied mask. Note that the *alien head* is clearly the object at that location as the alien head mask provides the smallest value of d.

Examination of neighborhood of pixel $(183, 57), d = 0.0352007$ without masking			
x coordinate	y coordinate	minimum eigenspace distance (d)	masking image
183	57	0.0129772	oralb bottle
183	57	0.0160544	porcelain cat
183	57	0.023429	wooden mushroom
183	57	0.0113579	plastic bottle
183	57	0.0150808	plastic cat
183	57	0.000325112	rolaid bottle
183	57	0.00799419	alien head
183	57	0.0721633	tomato
202	61	0.0657223	teacup

FIGURE 4.20. This table shows the minimum distances found in the 64×64 neighborhood of pixel $(183, 57)$ for each applied mask. Note that the *rolaid bottle* is clearly the object at that location as the rolaid bottle mask provides the smallest value of d.

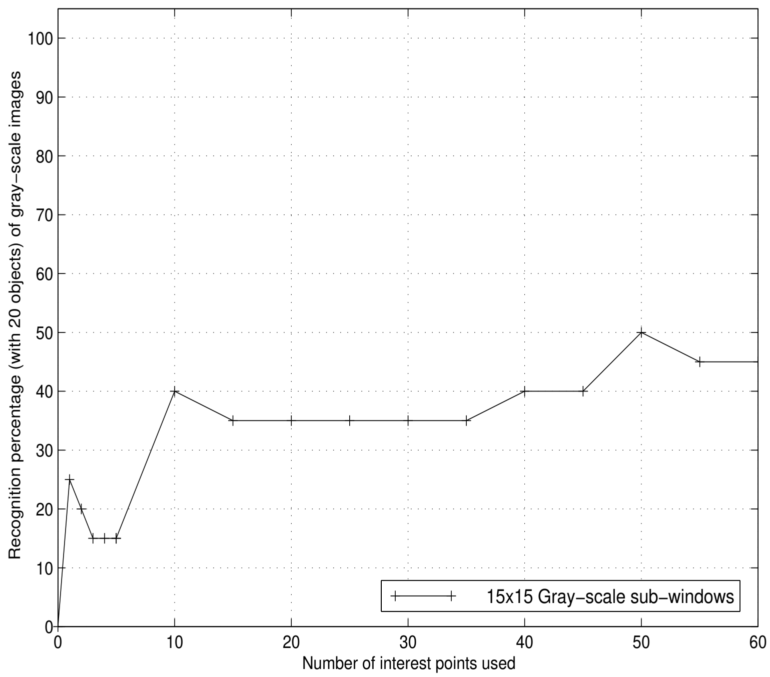
Examination of neighborhood of pixel $(194, 186), d = 0.0460267$ without masking			
x coordinate	y coordinate	minimum eigenspace distance (d)	masking image
194	186	0.0228138	oralb bottle
194	186	0.0240024	porcelain cat
194	186	0.031469	wooden mushroom
194	186	0.0219507	plastic bottle
194	186	0.000539501	plastic cat
194	186	0.0100043	rolaid bottle
194	186	0.0130058	alien head
213	184	0.0979818	tomato
214	186	0.0628941	teacup

FIGURE 4.21. This table shows the minimum distances found in the 64×64 neighborhood of pixel $(194, 186)$ for each applied mask. Note that the *plastic cat* is clearly the object at that location as the plastic cat mask provides the smallest value of d.

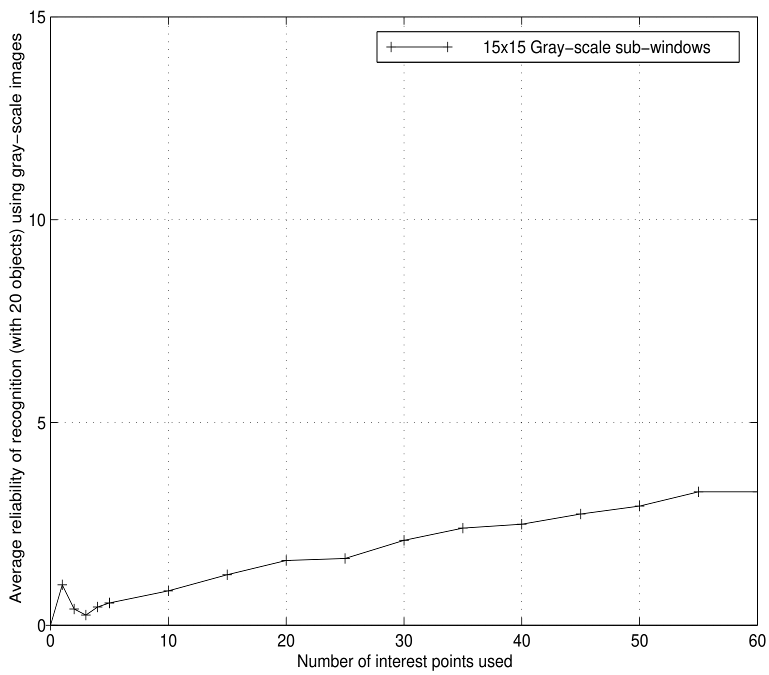
d	porcelain cat	rolaid bottle	alien head	tomato	plastic cat
Figure 4.14(a)	0.00121013	0.000325112	0.00209664	0.000329012	0.000539501
Figure 4.14(b)	0.00751744	0.00292928	0.0098226	0.00221511	0.00301337
Figure 4.14(c)	0.0015848	0.000496949	0.00285528	0.000442019	0.000726603
Figure 4.14(d)	0.00325844	0.00110525	0.00544933	0.00111467	0.00160119

FIGURE 4.22. This table shows the lowest values of d found in all the test images. They were all found at the appropriate locations and all the objects were correctly classified. See Figures 4.17 to 4.21 to see how the first row was constructed. The other rows were constructed similarly, except their tables were not included for reasons of brevity.

This can be explained by the fact that color pixels contain three times more information than their gray-scale counterparts. When using color we consider the amplitude spectra returned from the 2D Fourier transform of the images for each of the three colors (red, green and blue) in the RGB color space. We then combine this information when constructing the covariance matrix $\mathbf{Q}$ to obtain the eigenspace. In the gray-scale equivalent case one third of the information is lost as the RGB values are combined prior to the Fourier transform. The amplitude spectra thus returned is much less informative and consequently the construction of the eigenspace using this information is insufficient for reliable recognition.



(a) Recognition rate as a percent vs number of interest points



(b) Reliability of recognition rate vs number of interest points

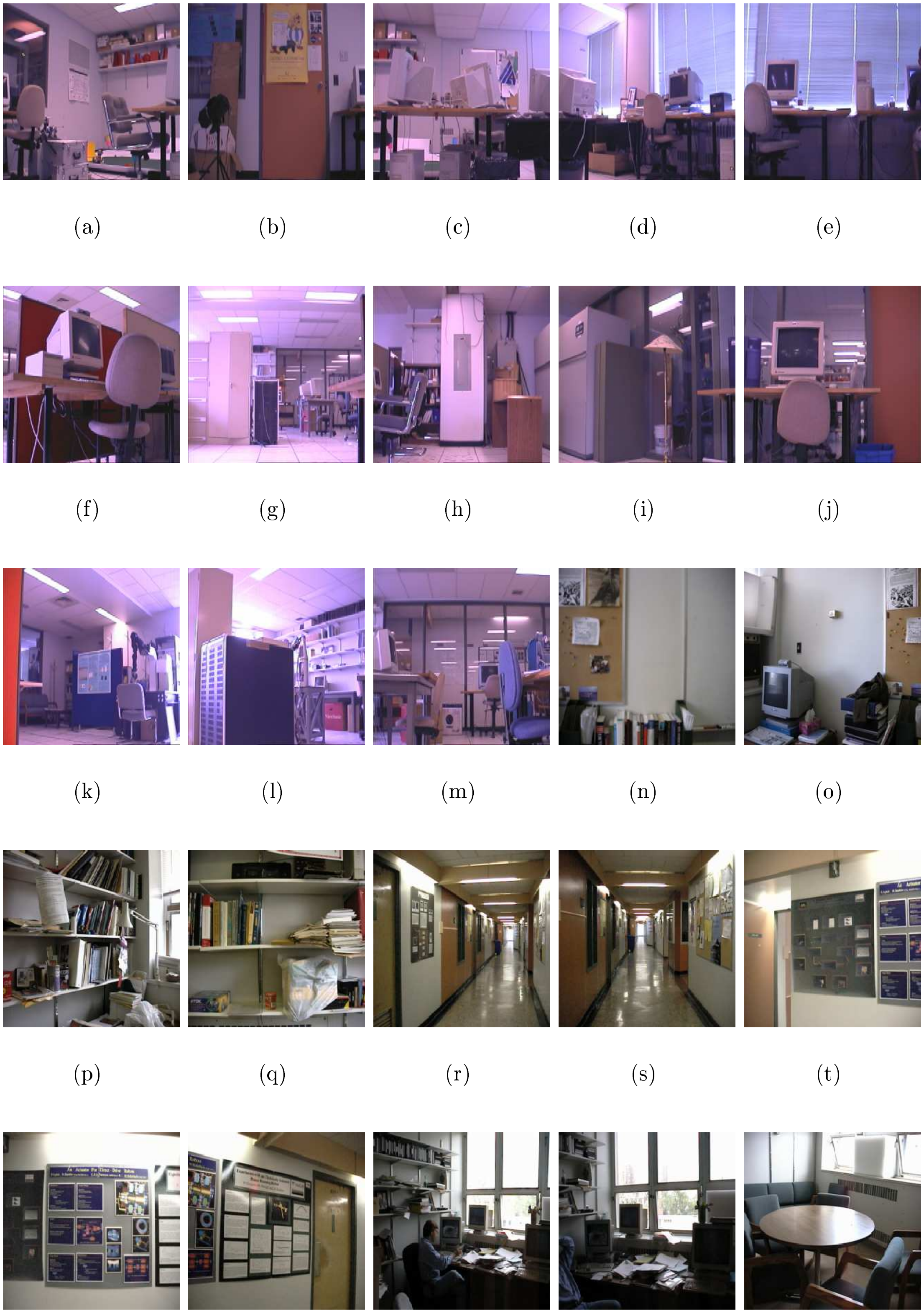
FIGURE 4.23. Results obtained when varying the number of interest points on *gray-scale* versions of the object images shown in Figure 4.5. The first graph shows the recognition performance on all the 20 images in Figure 4.5 as a percentage. The second graph shows the reliability of the recognition performed. See page 39 for a definition of reliability.

4.2. Place recognition

The place recognition problem provided better results than those obtained for the object recognition problem. This was entirely expected as we did not have to account for the background in this problem. While the background information is considered as noise in the object recognition experiments, here it is an inherent part of the room to be recognized thus no effort is required or made to account for it.

The eigenspace was created using the images shown in Figure 4.24 with 64 interest points. Sub-window sizes of 10x10 and 20x20 were used and as expected the results from the 20x20 case were more reliable. The tests themselves were performed using the 30 best eigenvectors and also 64 interest points. The test images we used can be seen in the first columns of Figures 4.25, 4.26, 4.27, and 4.28 with the 10x10 interest sub-windows chosen by the symmetry interest operator. While the results were more reliable using 20x20 (the definition of reliability remains unchanged from the previous section), all the test images were successfully classified using only 10x10 sub-windows. The corresponding best matches (selected from Figure 4.24) of all the test rooms are shown in the second columns of the same figures. Note that the test images are taken at different locations within the training rooms and that the camera alignment is different for these test cases. Note also that we demonstrate our systems ability to handle occluding objects as well as planar rotations of the image, see Figures 4.26(c), 4.26(e), 4.27(e). Figure 4.28(e) is interesting as it demonstrates our systems ability to handle view changes due to perspective transformations. The choice of small sub-windows where local perspective transformations are insignificant allows us to reliably recognize such images.

An interesting result which deserves special attention is the case for the recognition of Figure 4.25(e). While the best match found is shown adjacent to that image, the second best match found is also a legitimate classification. Figure 4.29 shows the test image followed by the two best matches found. The degree of overlap between the two training images is significant and that fact is captured by our recognition system in terms of the two best matches found for the given view.



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

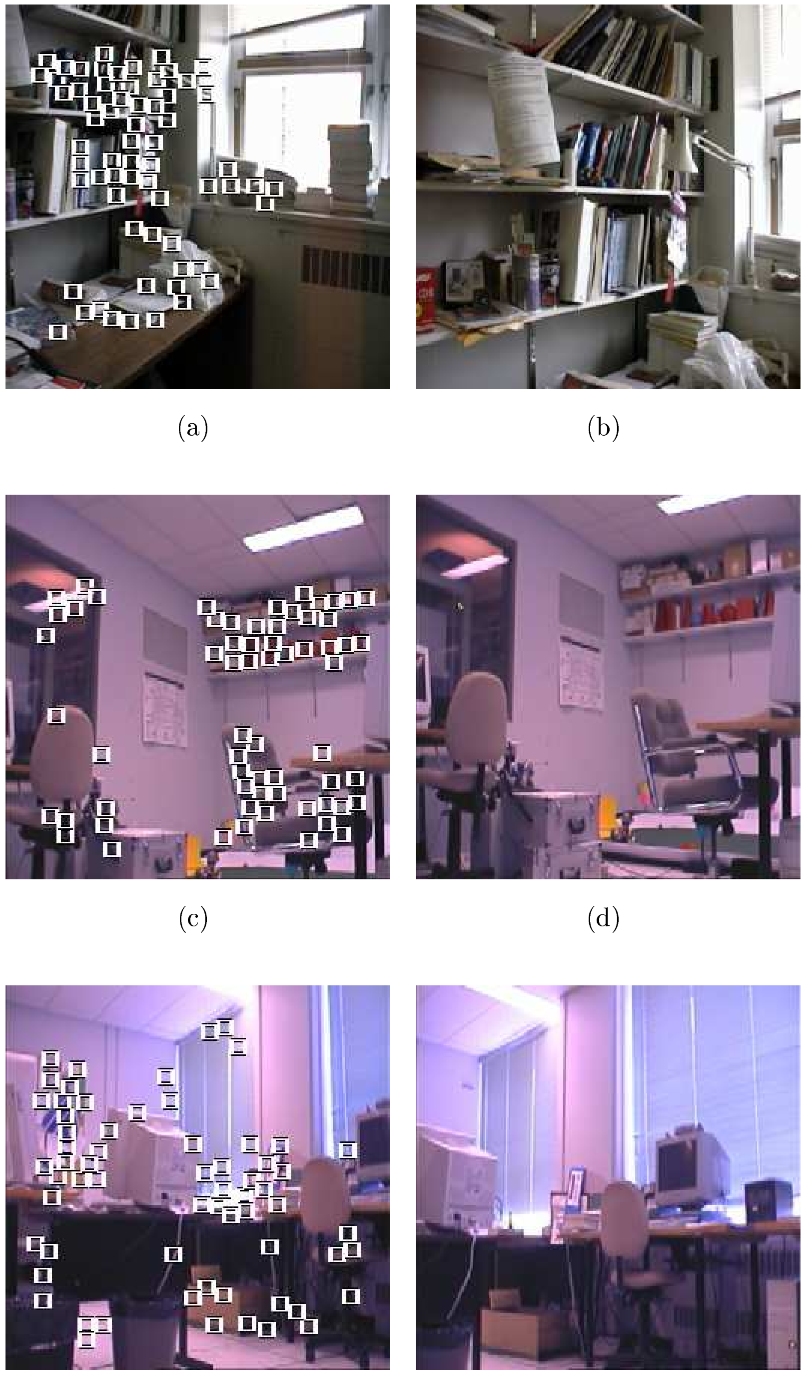
(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

(p) (q) (r) (s) (t)

(u) (v) (w) (x) (y)

FIGURE 4.24. Database of training rooms (all 256x256 images)



(a) (b) (c) (d) (e) (f)

FIGURE 4.25. Test images for room recognition (left column), followed by the best match (right column)

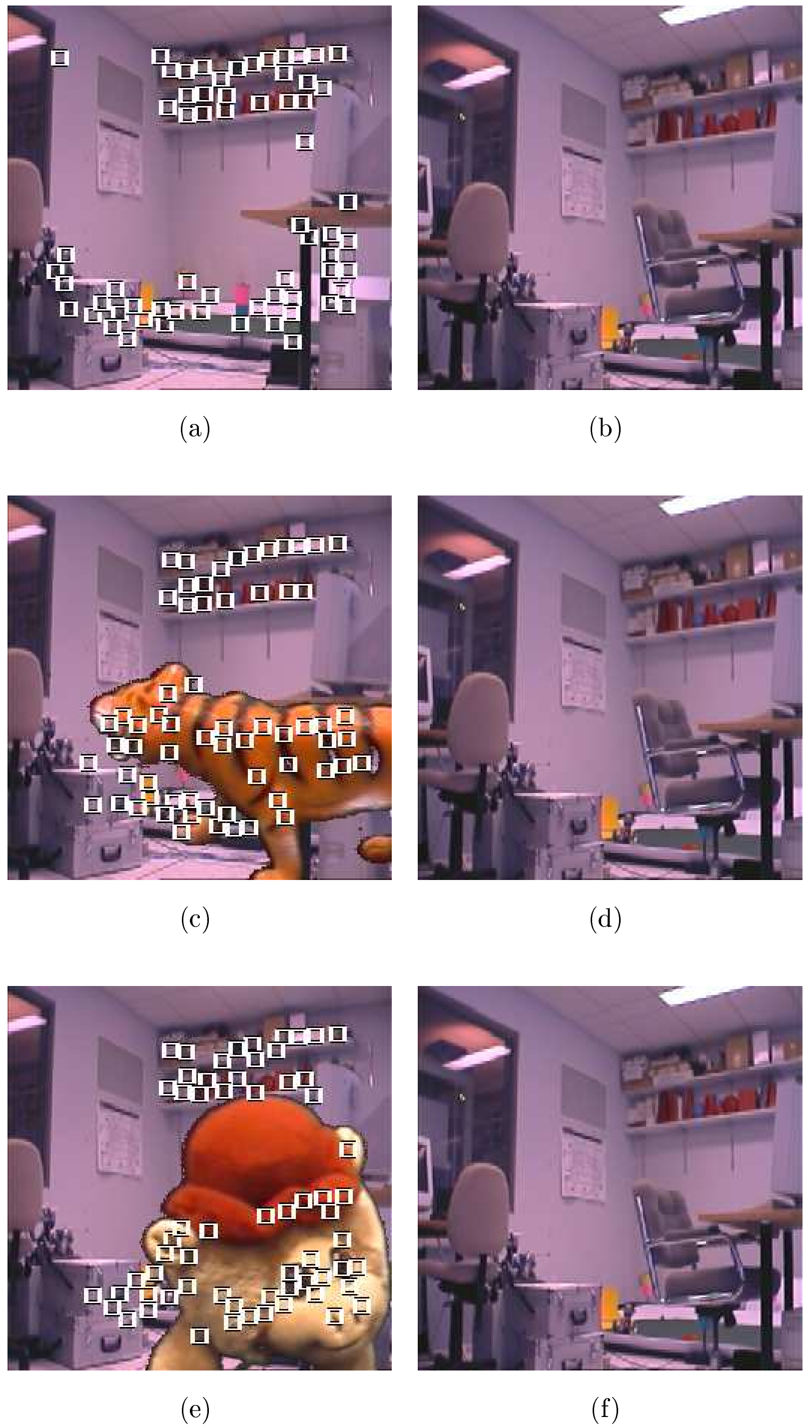


FIGURE 4.26. Test images for room recognition (left column), followed by the best match (right column)

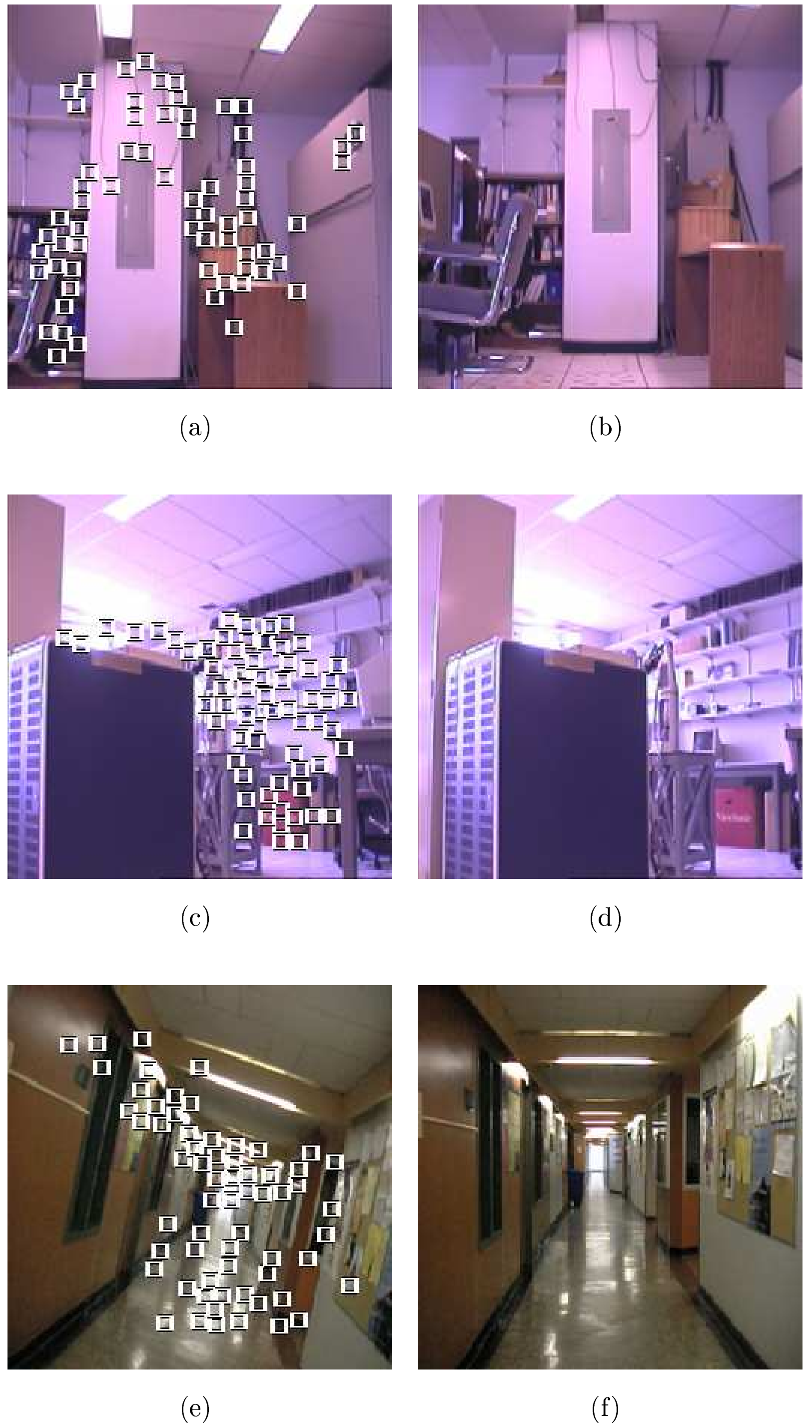
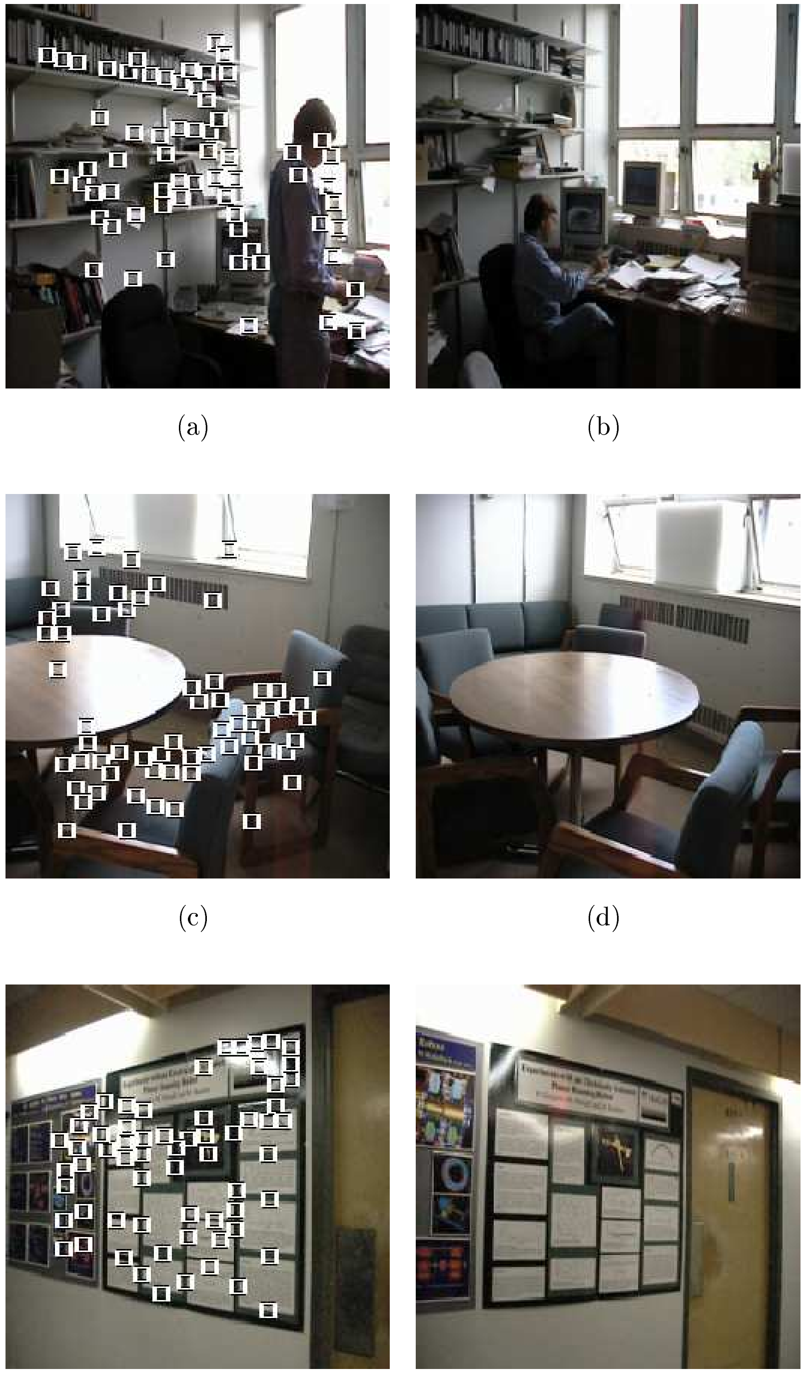


FIGURE 4.27. Test images for room recognition (left column), followed by the best match (right column)



(a) (b) (c) (d) (e) (f)

FIGURE 4.28. Test images for room recognition (left column), followed by the best match (right column)

(a) (b) (c)

FIGURE 4.29. Test image (a) and the two best matches returned. Note the overlap in images (b) and (c)

CHAPTER 5

Conclusions and Future Work

5.1. Conclusions

In this thesis we presented an approach to appearance based object recognition using single camera images. We first used an attention mechanism to select robust and informative interest points from an image. Sub-windows were then cropped around these features, filtered depending on their information content and consequently sampled using a polar mapping. The amplitude spectra of the frequency components of the resultant samples were then classified by their Euclidean distances to their nearest neighbors in a lower dimensional eigenspace constructed using PCA.

Our experiments showed that we were able to achieve a recognition rate of 100% with twenty 640x480 test images of fourteen distinct objects by only considering less than 5% of the content within the test images. The test images themselves consisted of, at times, planarly rotated, sometimes partially occluded objects placed on real non-uniform backgrounds. With the use of a radial symmetry attention operator we developed a robust PCA based object recognition system that successfully recognized objects in color under a wide variety of noisy conditions. The appropriately tuned attention mechanism allowed for the translations of the object under analysis as well as the selection of appropriate features for recognition. Rotation invariance was achieved due to the use of the amplitude spectra of the sub-windows sampled in a polar format. Our proposed masking technique accounted for the appearance of noisy background pixels in our sub-windows and the use of multiple features for classification allowed for partial occlusions of the objects from the camera.

We also conducted very successful experiments using the same approach for the recognition of rooms and places within our lab, correctly identifying 100% of the test images. Coarse localization using single camera images was performed successfully with the implemented system.

While far from solving the object recognition problem, we implemented and presented a promising technique which demonstrated the property that a few well selected features from the images are sufficient for both a wide variety of object and place recognition tasks once the features themselves have been processed in the ways we proposed. One does not necessarily need to segment an object from its background to successfully recognize it.

5.2. Future work

Many improvements can be made to our recognition system. First and foremost, handling the scale problem would be most beneficial. Creating multi-scale SIFT like keys [**11**] (see Section 2.2) with our current selected features would provide an added degree of freedom, allowing for the handling of a wide range of problems encountered in the object recognition field due to the scale problem.

Our masking technique is currently restricted to the spatial domain and thus fails when objects are rotated on the plane. We are currently working on a means of performing masking in the frequency domain by examining the frequency components of known background pixels in the training images.

To improve the robustness of our classifier, Schmid [**20**] proposes the idea of using the spatial relations between extracted features to improve classifications. Incorporating this information as a layer in our classification scheme could benefit our system and is part of our on-going research.

Similar sub-windows across different images or within the same image cluster together when projected onto a constructed eigenspace. While this property is at the heart of our classifier, problems arise as one ideally wants the selection of globally unique sub-windows from the images. Ohba and Ikeuchi [**16**] propose the elimination of training sub-windows by computing their similarity in eigenspace and only selecting ones which are greater that a predetermined threshold. The elimination of similar eigenwindows reduces the size of the training set while also making the classifier more stable as the classifier would no longer consider Euclidean distances from a large number of similar sub-windows but instead, have a distinct nearest neighbor.

Our current recognition model is restricted to the use of one interest operator per run. Ultimately we wish to implement a recognition model which simultaneously runs an assortment of interest operators on our images, independently processing the extracted features from them, and voting for an object based on the cumulative information from the various eigenspaces.

REFERENCES

[1] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman, *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection*, European Conference on Coputer Vision (Department of Electrical Engineering, Yale University, New Haven, CT 06520-8267), 1996, pp. 45–58.

[2] Irving Biederman, *Recognition-by-components: A theory of human image understanding*, Psychological Review **94** (1987), no. 2, 115–147.

[3] Eric Bourque, *On the automated construction of view points*, Master's thesis, McGill University, 1998.

[4] Eric Bourque and Gregory Dudek, *Viewpoint Selection – An autonomous robotic system for virtual environment creation*, vol. 1, October 1998, pp. 526–532.

[5] Virginio Cantoni, Stefano Levialdi, and Vito Roberto, *Artificial vision: Image description, recognition and communication*, Academic Press, 1997.

[6] Kenneth R. Castleman, *Digital image processing*, ch. 10, Prentice-Hall, 1996.

[7] Duda and Hart, *Pattern classification and scene analysis*, Wiley, New York. NY, 1973.

[8] Oja E., *Subspace methods of pattern recognition*, Research Studies Press, 1983.

[9] Murase H., Kimura F., Yoshimura M., and Miyake Y, *An improvement of the auto-correlation matrix in pattern matching methods and its application to handprinted hiragana*, Trans. IECE (1981).

[10] Jolliffe I.T., *Principal component analysis*, Springer-Verlag, 1986.

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

[12] David Marr, *Vision*, W. H. Freeman and Co., New York, 1982.

[13] H. Murakami and V. Kumar, *Efficient calculation of primary images from a set of images*, IEEE Trans. Pattern Analysis and Machine Intelligence **4** (1982), no. 5, 511–515.

[14] Hiroshi Murase and Shree K. Nayar, *Visual learning and recognition of 3-d objects from appearance*, International Journal of Computer Vision **14** (1995), 5–24.

[15] Shree K. Nayar, Hiroshi Murase, and Sameer A. Nene, *Learning, positioning, and tracking visual appearance*, Proc. IEEE Conference of Robotics and Automation (San Diego, CA), IEEE Press, May 1994, pp. 3237–3244.

[16] Kohtaro Ohba and Katsushi Ikeuchi, *Detectability, uniqueness, and reliability of eigen-windows for robust recognition of partially occluded objects*, IEEE Pattern Analysis and Machine Intelligence **19** (1997), no. 9, 1043–1048.

[17] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical recipes in c : The art of scientific computing*, Cambridge University Press, 1993.

[18] Daniel Reisfeld, Haim Wolfson, and Yehezkel Yeshurun, *Context free attentional operators: the generalized symmetry transform*, International Journal Of Computer Vision **14** (1995), 119–130.

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

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

[21] Cordelia Schmid, Roger Mohr, and Christian Bauckhage, *Comparing and evaluating interest points*, International Conference on Computer Vision, January 1998.

[22] Jianbo Shi and Carlo Tomasi, *Good features to track*, Computer Vision and Pattern Recognition, June 1994.

[23] Robert Sim and Gregory Dudek, *Robot positioning using learned landmarks*, McGill TR CIM-98-1170 (1998).

[24] ______, *Learning environmental features for pose estimation*, Proc. IEEE Workshop on Perception for Mobile Agents (Ft. Collins, CO), IEEE Press, June 1999.

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

[26] Carlo Tomasi and Takeo Kanade, *Detection and tracking of point features*, Tech. Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[27] Anne Triesman, *Perceptual grouping and attention in visual search for features and objects*, Journal of Experimental Psychology: Human Perception and Performance **8** (1982), no. 2, 194–214.

[28] ______, *Features and objects in visual processing*, Scientific American, November 1986.

[29] M. Turk and S. Pentland, *Eigenfaces for recognition*, Journal of Cognitive Neuroscience **3** (1991), 71–86.

[30] Matthew Turk and Alex Pentland, *Eigenfaces for recognition*, Journal of Cognitive Neuroscience **3** (1991), no. 1, 71–86.

Document Log:

Manuscript Version 1 —

Typeset by $\mathcal{AMS}$-LaTeX — 9 October 2000

DEEPTIMAN JUGESSUR

DEPARTMENT OF COMPUTER SCIENCE, 3480 UNIVERSITY STREET, MCGILL UNIVERSITY, MONTRÉAL, QUÉBEC H3A 2A7 CANADA,

E-mail address: `deepu@cim.mcgill.ca`