# Information Sampling for Appearance based 3D Object Recognition and Pose Estimation

**Niall Winters**[1,2],
[1]Computer Vision and Robotics Group,
Department of Computer Science,
University of Dublin, Trinity College,
Dublin 2 - Ireland.
Niall.Winters@cs.tcd.ie

**José Santos-Victor**[2],
[2]Instituto de Sistemas e Robótica,
Instituto Superior Técnico,
Av. Rovisco Pais, 1,
1049-001 Lisboa - Portugal.
jasv@isr.ist.utl.pt

**Abstract**

This paper is concerned with overcoming three problems associated with appearance based matching. The first is partial occlusion; the second background variation and the third is determining which image points - from a set of images acquired a priori - are the most discriminating. These data, either a single point or a number scattered throughout an image, are extracted by applying a statistical method we term *Information Sampling*.

We show how to use the data yielded by *Information Sampling* to build *Informative Local Appearance Spaces*. Preliminary results indicate that our method achieves successful object recognition and pose estimation while overcoming the difficulties outlined above.

## 1 Introduction

Since the early days of machine vision, object recognition has been a fruitful area of research for computer vision practitioners. Early systems [Besl and Jain, 1985] relied upon the geometric modeling of objects, a time consuming task. As an alternative to modeling, one can choose to remain in the image domain. Here, object recognition is viewed as a pattern recognition task, or more popularly as a problem which can be solved by using an *appearance based* solution.

Often, including our case, in order to compress large amounts of data, appearance based systems are built using Principal Component Analysis. Construction of such a system involves computing the eigenvectors (sometimes called eigenimages) of an a priori set of images. The variance of this set is captured by its first few eigenvectors. This low dimensional subspace [Murase and Nayar, 1995], also known as an eigenspace, forms an orthonormal basis into which each image from the a priori set is projected. Once this eigenspace has been built, real time recognition of an unknown image is achieved by projecting it into the eigenspace and using a simple distance measure to find its closest match to the previously projected points.

### 1.1 Related Research

Research into appearance based recognition has spawned a number of successful applications ranging from face recognition [Sirovich and Kirby, 1987, Turk and Pentland, 1991] to mobile robot navigation [Gaspar et al., 2000, Winters et al., 2000, Winters and Santos-Victor, 1999].

However, eigenspace matching is not the only instance of an appearance based method. Other approaches include colour histograms [Swain and Ballard, 1991], colour profiling [Duffy et al., 2000] and receptive field histograms [Schiele and Crowley, 1996].

By applying the appearance based paradigm, using eigenspace matching, to the area of object recognition, Murase and Nayar [Murase and Nayar, 1995] addressed the problem of automatically learning object models (thus avoiding geometric approaches) not only for recognition but also for pose estimation. Successful results were achieved by using a "global approach" to solve the problem, i.e. entire images were used for projection into the eigenspace. Their image set did not contain such aberrations as background variation, occlusion or scale change. It is well known that appearance-based methods have difficulty dealing with such adverse conditions. One of the issues addressed in this paper is how to attempt to deal with two of these aberrations, namely background variation and partial occlusion.

Overcoming partial occlusion has been the focus of a number of research works. Ohba and Ikeuchi [Ohba and Ikeuchi, 1997] divided the entire image into a number of subwindows, which they termed *eigenwindows*. Their basic premise was that even if a number of these eigenwindows were occluded, the remaining windows would contain enough information to identify an object. They did not deal with background variation.

*Uniform* background change (i.e. changing all of the background pixels to the same gray-level) was addressed by [Huttenlocher et al., 1996] when using an eigenspace approximation to the Hausdorff fraction. Unfortunately, they did not address the pose estimation problem.

### 1.1.1 Discriminatory Information Determination

A second point addressed in this paper is the extraction of the most *effective* information from a set of images. A method of determining the discriminating power of an eigenwindow was identified by Colin de Verdière and Crowley [de Verdière and Crowley, 1998]. Here, at the training stage, *every* window from every image was projected into the eigenspace. Naturally, all non-discriminating windows generated a large number of matches. Thus, suppression of these redundant windows was undertaken. The major downfall of this approach is that enough space and computational power is required to store and search all of the eigenwindows. Eliminating redundant windows was addressed by Ohba and Ikeuchi [Ohba and Ikeuchi, 1997] by utilising three criteria, namely: detectability, uniqueness and reliability.

An alternative to the eigenwindow approach is to search entire images for partially invariant image features, such as edge groupings, for example. Unfortunately, such features are often not detected frequently enough to allow for reliable recognition rates. Thus, for this method to exhibit improved reliability local features, where the signal change two-dimensionally (also known as interest points) are required. These can be determined by using an appropriate interest operator such as a Harris detector. Schmid and Mohr [Schmid and Mohr, 1997] used this approach for image retrieval from a large database.

One can combine the above two approaches, i.e. only use eigenwindows that contain a number of interest points above a certain threshold. This is the approach taken by Jugessur and Dudek [Jugessur and Dudek, 2000]. This approach requires highly textured images to function effectively. Our approach does not exhibit such a constraint and has been shown to work for images of low texture [Winters and Santos-Victor, 2001].

## 1.2 Our Approach

Our approach to the object recognition problem, utilises the inherent information contained within the image set. Essentially, our method termed *Information Sampling* minimizes (in some sense) the error covariance matrix associated with the reconstruction of an image from the object set using only a small number of (noisy) pixels. This is based on a method by Rendas and Perrone [Rendas and Perrone, 2000], although our method does not require the building of an eigenspace to determine the inherent information. Theoretically, it can be applied on a *pixel-by-pixel basis to any type of image*, as outlined in Section 2.1. In this paper, for computational

reasons, we use windows instead of pixels. We term these windows *Information Windows*. Once we have found each of these windows we rank them from most to least discriminatory. It is only *after* this stage that we use each of the information windows as the basis for building *Informative Local Appearance Spaces*, as outlined in Section 3. These are then used for object recognition and pose estimation.

This paper is outlined as follows: in Section 2 we detail the *Information Sampling* method and in Section 3 we present *Informative Local Appearance Spaces*. In Section 4 we give our experimental results and before drawing our conclusions and presenting the future directions of our research in Section 5.

## 2 The Information Sampling Method

As previously noted, our approach requires the use of a priori image data. We determined which regions contained the most relevant information, i.e. which were the most discriminatory by applying *Information Sampling*. As a first step in explaining this process, Section 2.1 outlines the procedure for reconstructing an image, given only a small amount of data.

### 2.1 Image Reconstruction

We assume that the our images can be modeled as a random vector $I$, characterized by a Gaussian distribution with mean $\bar{I}$ and covariance $\Sigma_I$:

$$I \sim \mathcal{N}(\bar{I}, \Sigma_I) = p(I)$$

Usually, one can take an ensemble of images $[I_1 \dots I_m]$, which can be utilized for computing $\bar{I}$ and $\Sigma_I$, so that $p(I)$ can be computed *a priori*. We assume that the observations, $d$, consist of a selection of (noisy) image pixels (or sub-regions), rather than the entire image. Accordingly, the observation model can be expressed as:

$$d = SI + \eta \tag{1}$$

where $d$ stands for the observed data and the measurement noise, $\eta$ is assumed to follow a Gaussian distribution with zero mean and covariance, $\Sigma_n$. We further assume that $I$ and $\eta$ are independent. The selection matrix, $S$ is composed of a series of ones and zeros, the ones corresponding to the data points extracted from an image. We select a number of pixels to test by moving the set of ones in the selection matrix.

Having prior knowledge of $I$, in the form of a statistical distribution, $p(I)$, the problem now consists of estimating the (entire) image based on partial (noisy) observations of a few pixels, $d$. This problem can be formulated as a *Maximum a Posteriori* estimation of $I$. The posterior probability can be determined from Bayes rule as follows:

$$p(I|d) = \frac{p(d|I)p(I)}{p(d)} \tag{2}$$

where $p(d|I)$ is the likelihood of a pixel (or set of pixels) given a known image, $I$; the prior distribution is denoted by $p(I)$ and is assumed to have been learnt *a priori*. With this information we calculate the maximum a posteriori estimate of an image, $\hat{I}_{MAP}$ as follows:

$$\hat{I}_{MAP} = \arg\max_I p(I|d) = (\Sigma_I^{-1} + S^T\Sigma_n^{-1}S)^{-1}(\Sigma_I^{-1}\bar{I} + S^T\Sigma_n^{-1}d) \tag{3}$$

Thus, $\hat{I}_{MAP}$ is the reconstructed image obtained using the pixel (or set of pixels), $d$. Notice that by combining the prior image distribution with the statistical observation model, we can estimate the entire image based on the observation of a *limited* number of pixels.

### 2.2 Choosing the Best Data: Information Windows

Once we have reconstructed an image using the selected data, we can compute the error associated with this reconstruction. The error covariance matrix, $\Sigma_{error}$ is given by:

$$\Sigma_{error} = \text{Cov}(I - \hat{I}_{MAP}) = (\Sigma_I^{-1} + S^T \Sigma_n^{-1} S)^{-1} \tag{4}$$

Of course, the quality of the estimate, and the "size" of $\Sigma_{error}$ depend not only on the observation noise, $\eta$ but also on the observed image pixels, as described by the selection matrix, $S$. Equation (4) quantifies the quality of an estimate obtained from using a particular set of image pixels. In theory, we can evaluate the *information content* of any individual image pixel or combination of pixels, simply by selecting an appropriate selection matrix, $S$, and determining the associated $\Sigma_{error}$.

This problem could be formulated as an experiment design process, in which we look for the optimal selection matrix $S^*$ that minimizes (in some sense) the error covariance matrix. If we take the determinant of $\Sigma_{error}$ as an indication of the "size" of the error, the optimal selection of image pixels would be given by:

$$S^* = \arg\min_S \{\ det((\Sigma_I^{-1} + S^T \Sigma_n^{-1} S)^{-1})\ \} \tag{5}$$

In practice, to avoid computing the inverse we define the following equivalent optimization problem in terms of a modified uncertainty metric, $U$:

$$U = -\log\{\ det(\Sigma_I^{-1} + S^T \Sigma_n^{-1} S)\ \}\ ; \quad S^* = \arg\min_S U \tag{6}$$

So far, we have described *Information Sampling* as a process for (i) reconstructing an entire image from the observation of a few (noisy) pixels and (ii) determining the *most relevant* image pixels, $S^*$, in the sense that they convey the most information about the image set.

Unfortunately, determining $S^*$ is computationally impractical since we would have to compute $\Sigma_{error}$ for all possible combinations of pixels scattered throughout the image. Instead, we partition the image into non-overlapping square windows of $(l \times l)$ pixels. We term these regions *Information Windows*, denoted by $\mathbf{w} = [w_1 \dots w_n]$.

By using equation (6), we can rank *Information Windows* or combinations of such windows, in terms of their information content. Again, as searching for all possible combinations of windows within the image, in order to minimize equation (6), would be computationally intensive, we instead use two sub-optimal (greedy) algorithms. For details on each, see [Winters and Santos-Victor, 2001].

Notice that the information criterion is based on the entire set of images and not, as with other methods, on an image-by-image basis. For instance, a highly textured image region would only be selected if it varied significantly from one image to the next.

## 3 Informative Local Appearance Spaces

Given the information windows, we extract a region from all images in the object set, $\mathbf{I}^w$ corresponding to each window, $\mathbf{w} = [w_1 \dots w_n]$. Following this step we apply PCA to build a local appearance space using *only* $\mathbf{I}^w$. This space we term an *Informative Local Appearance Space (ILAS)*. Each information window has an associated ILAS, the higher the ILAS ranking more discriminating power it exhibits. The ILAS eigenvectors are of length $l^2$, where $l$ is the length of an information window. Object recognition begins using the highest ranking ILAS. It is only when a significant part of an information window is occluded that the next most discriminating ILAS is loaded, etc. Implicitly, this allows use to deal with partial occlusion.

The advantage of this approach over the related research outlined in Section 1 is that we immediately know which information window from each image gives us the highest recognition rate. Thus, we need only build a local eigenspace using each information window. In essence, we extract quality information rather than relying on a large quantity of information.

# 4 Experimental Results

Object recognition and pose estimation experiments were performed on a 450 MHz Pentium III PC using Matlab. The object set used was the COIL-20 database [Nene et al., 1996], as shown in Figure 1. Each image is 128 × 128 pixels in size. Experiments were undertaken using 36


Figure 1: A selection of images from the COIL-20 database.

evenly spaced views of 20 objects as the database set and a different 36 evenly spaced views of the same 20 objects as the test set.

## 4.1 Finding and Ranking Information Windows

We ran our *Information Sampling* method on the COIL-20 database, to determine the most discriminatory information windows. Due to computational constraints, each 128 × 128 image was first subsampled to 32 × 32 pixels in size. The reason for such an image size relates to the complexity of determining the error covariance matrix, $\Sigma_{error}$ in equation (4). Each information window was chosen to be 8×8 pixels in size, thus giving 16 non-overlapping information windows per image, ordered from left-to-right and top-to-bottom. Once the information windows were ranked, corresponding 32 × 32 information windows in the 128 × 128 sized images were found using the simple ratio: 64:1,024=1,024:16,384. These information windows were used to build each ILAS.

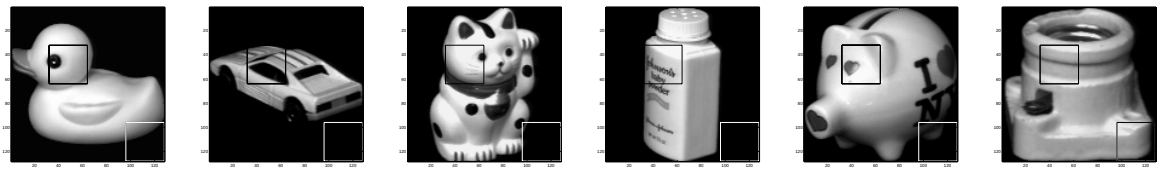

Figure 2: A selection of images showing the highest (mid-image) and lowest ranking (bottom-right) information windows, respectively in a selection of images.

Figure 2 shows six objects from the database in a number of differing poses along with their associated most (shown mid-image) and least (shown at the bottom-right of the images) discriminating information windows, as yielded by the *Information Sampling* method. Notice that each information window discriminates over the entire set of images, not on an image-by-image basis.

### 4.2 Matching Results

Object recognition and pose estimation experiments were first undertaken on unperturbed images using only ILAS 1, i.e. the highest ranking appearance space. This is an improvement over previous approaches, where *all* windows first had to be projected into a local eigenspace before recognition could occur. Thus, we were able to immediately reduce the ambiguity associated with projection. In addition to the compression yielded by PCA, further compression to one sixteenth of the original image size was achieved by using information windows.

The images in Figure 3 show the results obtained using the highest ranking information window. In both cases, the left image shows this window, extracted from the *unknown* object we wish to recognise (middle), and the right image shows the closest match at the correct pose. Results obtained using a large set of 720 unknown images reveal that the correct object was determined in 95.3% of cases and the correct pose in 73.8% of cases. Importantly, if other

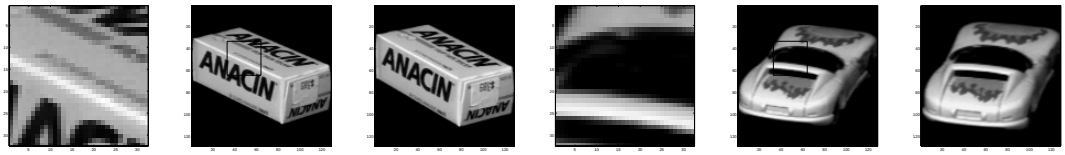


Figure 3: Object recognition and pose estimation without background variation. When using the *most* discriminating information window, the correct recognition rate is 95.3% and the correct pose estimation rate 73.8%.

regions of the image were occluded but the information window used for recognition was not, then the recognition results did not deteriorate. On the other hand, if an information window was occluded then the method outlined in Section 4.3 was used to overcome the problem. We obtain recognition results comparable to those using entire images for matching, but utilize significantly less image data.

In order to test the discriminating power of each information window we compared matching results using the $1^{st}$ and $3^{rd}$ most discriminating windows. In this case, ILAS 3 yielded a correct object recognition rate of 82.5% and a pose estimation rate of 65.3%. Thus, as expected the discriminating power of ILAS 1 is superior. Naturally, the lower the ILAS level, the more our approach degrades.

### 4.3 Results: Non-Uniform Background Change

As a further test of our method we decided to run it on images with *non-uniform* background variation. This is a particularly difficult problem, as PCA is well known to be susceptible to such changes. Since an information window may contain some background data or may be partially occluded, we wish to minimize the effect of such aberrations. Thus, we added an additional step to our method. Once we determined each information window, *we subdivided it into* 16 *subwindows*. These subwindows (and not the information windows) were then used to build each Informative Local Appearance Space. Background variation was dealt with by associating

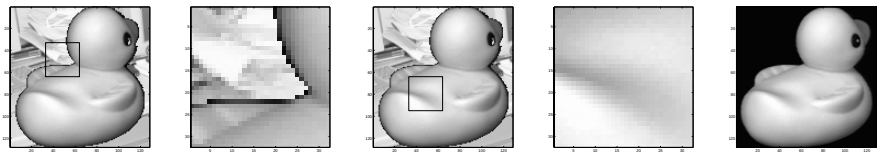

Figure 4: Object recognition and pose estimation with non-uniform background variation.

a confidence level to each information window. If a high percentage of the subwindows identify the same object we trust the result. If this is not the case, then most of the subwindows fall on the background region and not the object itself. Then, object recognition can be achieved by jumping to the next ILAS, and repeating the process. This is shown in Figure 4, where the

| Image Change | Correct | False Positive | No ID(6) |
|---|---|---|---|
| Unperturbed (ILAS 1) | 95.3% | 4.7% | - |
| Unperturbed (ILAS 3) | 82.5% | 17.5% | - |
| Non-Uniform Background Change | 87.6% | 5.5% | 6.9% |

Table 1: Object Recognition Results Summary.

| Image Change | Correct | False Positive |
|---|---|---|
| Unperturbed (ILAS 1) | 73.8% | 16.2% |
| Unperturbed (ILAS 3) | 65.3% | 34.7% |
| Non-Uniform Background Change | 50% | - |

Table 2: Pose Estimation Results Summary.

most informative information window is identified as containing a large amount of non-uniform background variation. In this case, object recognition and pose estimation were successfully achieved using ILAS 3. For information windows with less background variation, jumping to the next ILAS is not necessary. Using 612 test cases and the first six information windows, the correct object was identified in 87.6% of cases, with a false positive rate of 5.5%. An object was unidentifiable in 6.9% of cases. Tables 1 and 2 summarize the results obtained.

## 5 Conclusions and Future Work

This paper presented *Information Sampling* and *Informative Local Appearance Spaces (ILAS)* to improve appearance based matching. Preliminary results were detailed.

In the near term our future work shall be directed towards improving the method in a number of ways. Firstly, the best information windows per object, rather than that best windows over the entire set of objects shall be determined. In this way each information window would be tailored to each object. Increasing the number of information windows available shall be undertaken. Finally, robust statistics shall be integrated into the method.

## Acknowledgements

This work was partly funded by the European Union RTD - Future and Emerging Technologies Project Number: IST-1999-29017, Omniviews.

## References

[Besl and Jain, 1985] Besl, P. and Jain, R. (1985). Visual recognition using local appearance. *ACM Computing Surveys*, 17(1):75–145.

[de Verdière and Crowley, 1998] de Verdière, V. C. and Crowley, J. L. (1998). Visual recognition using local appearance. In *5th European Conference on Computer Vision, (ECCV 1998)*, pages 640–654, Freiburg, Germany.

[Duffy et al., 2000] Duffy, N., Crowley, J. L., and Lacey, G. (2000). Object detection using colour. In *15th International Conference on Pattern Recognition*, pages 640–654, Barcelona, Spain.

[Gaspar et al., 2000] Gaspar, J., Winters, N., and Santos-Victor, J. (2000). Vision-based navigation and environmental representations with an omni-directional camera. *IEEE Transactions on Robotics and Automation*, 16(6):890–898.

[Huttenlocher et al., 1996] Huttenlocher, D., Lilien, R., and Olson, C. (1996). Object recognition using subspace mathods. In *Proceedings of the 4th European Conference on Computer Vision*, pages 536–545, Cambridge, UK.

[Jugessur and Dudek, 2000] Jugessur, D. and Dudek, G. (2000). Local appearance for robust recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 834–839, Hilton Head Island, SC, USA.

[Murase and Nayar, 1995] Murase, H. and Nayar, S. K. (1995). Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24.

[Nene et al., 1996] Nene, S., Nayar, S., and Murase, H. (1996). Columbia object image library (coil-20). Technical Report CUCS-005-96, Columbia University.

[Ohba and Ikeuchi, 1997] Ohba, K. and Ikeuchi, K. (1997). Detectibility, uniqueness and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1043–1048.

[Rendas and Perrone, 2000] Rendas, M. and Perrone, M. (2000). Using field subspaces for online survey guidance. In *Proceedings of Oceans 2000*, Providence, RI, USA.

[Schiele and Crowley, 1996] Schiele, B. and Crowley, J. L. (1996). Object recognition using multidimentional receptive field histograms. In *4th European Conference on Computer Vision, (ECCV 1996)*, pages 610–619, Cambridge, UK.

[Schmid and Mohr, 1997] Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535.

[Sirovich and Kirby, 1987] Sirovich, L. and Kirby, M. (1987). Low dimentional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524.

[Swain and Ballard, 1991] Swain, M. and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

[Turk and Pentland, 1991] Turk, M. A. and Pentland, A. P. (1991). Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'91)*, pages 586–591.

[Winters et al., 2000] Winters, N., Gaspar, J., and Santos-Victor, J. (2000). Omni-directional vision for robot navigation. In *Proceedings of the 1st International IEEE Workshop on Omnidirectional Vision at CVPR 2000*, pages 21–28, Hilton Head Island, SC, USA.

[Winters and Santos-Victor, 1999] Winters, N. and Santos-Victor, J. (1999). Omni-directional visual navigation. In *Proceedings of the 7th International Symposium on Intelligent Robotics Systems*, pages 109–118, Coimbra, Portugal.

[Winters and Santos-Victor, 2001] Winters, N. and Santos-Victor, J. (2001). Information sampling for optimal image data selection. In *Proceedings of the 9th International Symposium on Intelligent Robotics Systems*, pages 249–257, Toulouse, France.