Appearance Matching with Partial Data *

Efstathios Hadjidemetriou and Shree K. Nayar
Department of Computer Science
Columbia University
New York, NY 10027

Abstract

Appearance matching methods use raw or filtered pixel brightness values to perform recognition. To expedite recognition, subspace methods are used to achieve compact representations of images. In many cases it is advantageous to recognize an image based on only a subset of its pixels, for example, when a part of an image is occluded, or to expedite recognition. Currently, such subsets are selected either randomly or using heuristics. In this paper, we derive criteria for selecting the pixel subsets through a sensitivity analysis of the subspace. Based on these criteria, we propose two practical recognition algorithms. These algorithms were tested on a large number of images with degraded or partial data. In addition to faster recognition, our algorithms yield high recognition accuracy.

1 Introduction

Appearance matching based on linear subspace methods have found many important applications in computational vision, including, face recognition [Turk and Pentland, 1991], real-time 3D object recognition [Nayar *et al.*, 1996], and planar pose measurement [Krumm, 1996]. Appearance matching methods generally use image brightness values directly, without relying on the extraction of low-level cues such as edges, local shading, and texture. The success of this approach results from the fact that brightness values capture both geometric and photometric properties of the objects of interest.

There are at least two reasons that motivate us to use a subset of the pixels in the image, rather than the complete image. First, if an image includes occlusion we would like to use only the incorrupted pixels for recognition. Secondly, using a subset of the pixels can enhance efficiency because recognition time in appearance matching is more or less proportional to the number of pixels used.

A number of attempts have been made to perform recognition with occluded or partial data [Murase and Nayar, 1995a] [Moghaddam and Pentland, 1995] [Krumm, 1996] [Brunelli and Messelodi, 1993] [Leonardis and Bischof, 1996]. Though these approaches are interesting, none succeeds to address the underlying problems fully. The first three techniques select windows in an image based on ad–hoc heuristics that are not generally applicable. The last two methods, at first, randomly select a small subset of pixels and then prune the subset with iterative algorithms. However, these iterative schemes are not guaranteed to converge to the desired recognition result. In addition, recognition based on very small random subsets is not generally reliable.

The more general problem of using partial data has been investigated thoroughly in the context of statistics [Gauss, 1873] [Hotelling, 1944] [Ehrenfeld, 1955] [Huber, 1981] [Cook and Weisberg, 1982] [Box and Draper, 1975]. These results are insightful but are limited in their applicability as they use assumptions that do not hold true in appearance matching. For instance, the data sets are assumed to be small (few pixels) and the measurements are assumed to be repeatable (multiple measurements at each pixel).

In this paper, we derive several criteria for selecting subsets of image pixels that maximize recognition rate. This is accomplished by analyzing the sensitivity of the subspace to image noise. Our criteria are then used to develop recognition algorithms that are general in their applicability. The first algorithm automatically selects a square window within an image as the pixel subset. The use of such a window reduces sensitivity of recognition to occlusion. This is due to the fact that occlusion is more likely to appear in a large image rather than a small window. The second algorithm judiciously selects the subset from the entire image, i.e. the pixels are not restricted to lie within a local region. Both algorithms are tested with a large number of noisy images. They demonstrate higher recognition performance when compared to algorithms that select pixel subsets randomly.

*This work was supported in part by ONR/DARPA MURI program under ONR Contract No. N00014-95-1-0601. Several other agencies and companies have also supported aspects of this research.

2 Overview of Appearance Matching

The traditional approach to appearance matching consists of two stages: model acquisition and recognition. During model acquisition, a training set is obtained by varying a number of parameters (object pose, illumination, etc). Each image in the training set is read in a raster scan fashion to yield an n–dimensional vector, which is then normalized to unit energy to achieve invariance with respect to illumination intensity. Next, the correlation matrix of the training images is constructed and its eigenvalues and eigenvectors are computed. Since the training images are normalized and the correlation matrix is symmetric, the eigenvectors are orthonormal. In general, a small number, k, of the eigenvectors (the ones with the largest eigenvalues) are sufficient to capture the primary variations in the training set. These eigenvectors form the basis vectors of a subspace called the eigenspace.

In the next step of model acquisition, each training image vector is projected to the subspace by computing the dot product of it with each of the basis vectors. The projections of all the training images yields a set of discrete points in the subspace. Images that are strongly correlated project to points that are close to each other. The discrete points can be interpolated [Murase and Nayar, 1995b] to get a manifold that represents all possible appearances of the object.

During recognition, each novel image is first normalized and then projected to the subspace as follows:

$$\widehat{\underline{x}} = A^T \underline{b} \tag{1}$$

where A is the orthonormal matrix whose columns are the eigenvectors of the training set, $\underline{b}$ is the normalized image vector, and $\widehat{\underline{x}}$ is the coordinate vector of the projection. Eventually, the closest manifold point to the projection $\widehat{\underline{x}}$ is found and the test image is identified as the one that corresponds to that point.

The projection $\widehat{\underline{x}}$ can also be used to reconstruct the image. This is done by substituting equation (1) into the linear relation between the eigenspace and the test image, namely $\widehat{\underline{b}} = A\widehat{\underline{x}}$, to obtain:

$$\widehat{\underline{b}} = AA^T \underline{b} = \mathbf{H}\underline{b} \tag{2}$$

where, $\mathbf{H}$ is the projection matrix. In turn, the reconstructed image can be used to calculate the fitting error, or residual image:

$$\underline{r} = \underline{b} - \widehat{\underline{b}} = (I - \mathbf{H})\underline{b} \tag{3}$$

where I is the $n \times n$ identity matrix, and n is the number of pixels in the image. Note that $\mathbf{H}$ is symmetric and idempotent; therefore, its diagonal elements satisfy $0 \leq h_{ii} \leq 1$ [Strang, 1980].

In the traditional approach to appearance matching, the complete set of all n image pixels is used. In our work, we seek to use a subset of m pixels, where $k < m \leq n$. That is, $\underline{b}$ is an m–dimensional vector whose elements are the intensity values that correspond to the pixel subset. In this context, A, which represents the subspace, consists of the subset of rows that corresponds to the pixel subset. In this case A is not necessarily orthonormal. Therefore, the orthogonal projection is obtained as:

$$\widehat{\underline{x}} = (A^T A)^{-1} A^T \underline{b} \tag{4}$$

Further, by substituting equation (4) into the linear relation $\widehat{\underline{b}} = A\widehat{\underline{x}}$, we get the projection matrix

$$\mathbf{H} = A(A^T A)^{-1} A^T \tag{5}$$

where $\mathbf{H}$ is $m \times m$ in size. Finally, $\widehat{\underline{b}}$ and r are also m–dimensional vectors.

The results reviewed above are used in the sensitivity analysis of design matrix A. For clarity, the design and projection matrices that correspond to the complete pixel set (full image) are denoted as A^c and $\mathbf{H^c}$, respectively.

3 Sensitivity Analysis of the Rows of the Design Matrix

In many real world applications the test image is corrupted by noise. In turn, the noise degrades the estimates of the subspace projection $\widehat{\underline{x}}$ and the reconstructed image $\widehat{\underline{b}}$. In addition, the degradation of $\widehat{\underline{x}}$ and $\widehat{\underline{b}}$ also depends on the properties of the rows of A. In this section we derive the properties that the rows of A should satisfy in order to minimize the degradation due to noise.

We assume that the noise degrading the test image is additive, independently distributed, has zero mean, and has finite variance. That is:

$$E(\underline{e}) = 0 \tag{6}$$

$$E(\underline{e}\underline{e}^T) = \sigma^2 I \tag{7}$$

where $\underline{e}$ is the random noise vector, and σ^2 is the variance of the noise. Hence, the image vector $\underline{b}$ can be decomposed into

$$\underline{b} = \underline{b}^u + \underline{e} \tag{8}$$

where $\underline{b}^u$ is the underlying incorrupted vector. The noise vector $\underline{e}$ is unobservable since part of it lies in the subspace. Hence, $\underline{b}^u$ and $\underline{e}$ cannot be recovered. However, they can be approximated by $\widehat{\underline{b}}$ and $\underline{r}$, respectively. If the approximation is close enough, then the reconstructed vector $\widehat{\underline{b}}$ is a reasonable estimate of incorrupted vector $\underline{b}^u$ and the residual $\underline{r}$ is a reasonable estimate of the unobservable noise $\underline{e}$.

A useful expression for the residual is obtained by substituting relation (8) into equation (3), and using $\underline{b}^u = \mathbf{H}\underline{b}^u$:

$$\underline{r} = (I - \mathbf{H})\underline{e} \tag{9}$$

The i^{th} row of matrix equation (9) gives the relation between the unknown noise and the residue in the i^{th} pixel [Cook and Weisberg, 1982]:

$$\begin{aligned} r_i &= e_i - \hat{e}_i \qquad (10) \\ &= e_i - \sum_{j=1}^{j=m} h_{ij} e_j = (1 - h_{ii}) e_i - \sum_{j=1, j\neq i}^{j=m} h_{ij} e_j \end{aligned}$$

where, e_i and $\hat{e}_i$ are, respectively, the actual noise and the part of the noise that lies in the subspace.

In equation (10), if the diagonal element h_{ii} and the non–diagonal ones h_{ij} are both small, then r_i will be a reasonable estimate for the unobservable noise e_i. The diagonal element can be made small by making it equal to its average value [Box and Draper, 1975]. The average value is given by $\frac{k}{n}$ since the sum of the diagonal elements of H is fixed,

$$\sum_{i=1}^{n} h_{ii} = k, \qquad (11)$$

where k is the rank of A.

To examine the magnitude of the non–diagonal elements we use their relation with the diagonal ones:

$$\sum_{j=1}^{n} h_{ij}^2 = h_{ii} \qquad (12)$$

This relation is derived using the fact that projection matrix $\mathbf{H}$ is idempotent and symmetric. By rearranging this expression we obtain:

$$h_{ii}(1 - h_{ii}) = \sum_{j=1, j\neq i}^{n} h_{ij}^2 \qquad (13)$$

This expression shows that the sum of the squares of the non–diagonal elements in the i^{th} row or column is a parabolic function of h_{ii}. This function is illustrated in Figure 1. It is clear from the figure that when h_{ii} is small, then the sum of the squares of the non–diagonal elements is also small. This holds when h_{ii} lies in the range $0 \leq h_{ii} \leq 0.5$, which is always the case since for a large number of pixels the average value of the diagonal elements, $\frac{k}{n}$, is a very small number, well below 0.5. Hence, by examining the conditions that keep both the diagonal and non–diagonal elements of the projection matrix small, we have shown that:

THEOREM 1 *The residuals will be close to the unknown noise if the diagonal elements of* $\mathbf{H}$ *are equal.*

This relates to a theorem of Huber [Huber, 1981] which says that the maximum diagonal element of $\mathbf{H}$ should tend to zero.

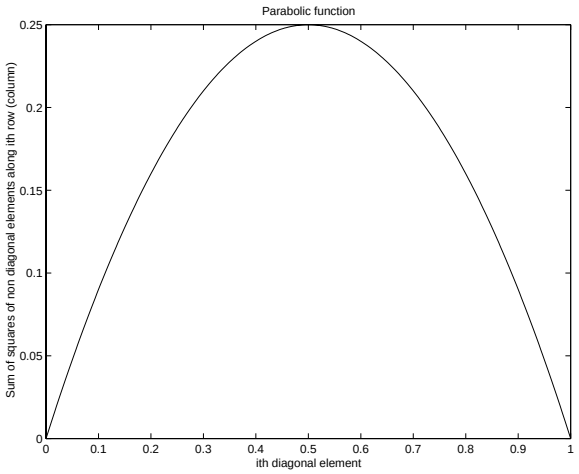


Figure 1: The sum of the squares of the non–diagonal elements of the projection matrix in the i^{th} row(column) as a function of the i^{th} diagonal element. Clearly, for small values of the diagonal element the sum of the squares of the non–diagonal elements is also small.

The above theorem leads to a useful geometrical implication, by regarding the rows of A as vectors in k–dimensional space. In particular, from the relation $H = A(A^T A)^{-1} A^T$ we can see that

$$h_{ii} = a_i (A^T A)^{-1} a_i^T. \qquad (14)$$

where a_i is the i^{th} row of A. The spectral decomposition of $A^T A$, being positive definite, gives eigenvalues $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k > 0$, with corresponding unit eigenvectors $p_1, p_2, \ldots, p_k$. By substituting the covariance matrix $A^T A$ with its spectral decomposition in (14) we get

$$h_{ii} = \sum_{l=1}^{k} (\frac{p_l^T a_i}{\sqrt{\mu_l}})^2 \qquad (15)$$

Further, letting θ_{li} denote the angle between p_l and a_i, we obtain

$$h_{ii} = a_i^T a_i \sum_{l=1}^{k} \frac{\cos^2 \theta_{li}}{\mu_l} \qquad (16)$$

Clearly, the magnitude of h_{ii} depends on two factors, namely $a_i^T a_i$, and the summation. Hence, we can make the diagonal elements comparable if we make both factors comparable. Considering the first we obtain:

- **Condition (1): The magnitudes of the rows of A should be comparable.**

The second factor is a summation of the projections of the inverse of the eigenvalues of all eigenvectors in the direction of row a_i. The summation of the projections must be comparable for all rows. In general, rows lie in almost all directions in space. Hence, the summation must be the same for all directions in space, that is, it

must be independent of orientation. In other words, we have:

- **Condition (2): The energy of the row vectors of A should be uniformly distributed in k-dimensional space.**

Note that the energy of a row is its euclidean length. The two conditions lead to the conclusion that, ideally, the rows of A should be uniformly distributed on the surface of a sphere.

There is no mathematical technique to find the optimal subset that most closely satisfies the two conditions. Hence, the optimal subset can only be found by examining all $\binom{n}{m}$ possible combinations of m pixels. The number of possible combinations is a very large number, hence, the computational complexity of finding the optimal subset is very high. However, the complexity of finding a suboptimal subset is significantly lower. A suboptimal subset can be found by techniques that approximate the conditions. As a first approximation we give priority to the rows that lie in under–represented directions in space, in order to satisfy condition (2). The rows are selected from the complete set of n rows, which gives rise to the projection matrix $H^c = A^{cT}A^c$. Hence, its diagonal elements are equal to $h^c_{ii} = a^{cT}_i a^c_i$. However, the diagonal elements are also given by (16). By comparing the two relations for h^c_{ii} we conclude that the factor that represents the summation of the energy in (16) is always equal to one. In turn, this implies that the row energy of an orthonormal matrix is uniformly distributed in space. To achieve the uniform distribution, the row vectors that lie in under–represented directions are compensated by having larger magnitudes. Hence, to give priority to under–represented directions we have:

- **Heuristic (1): Pixels that correspond to diagonal elements of $\mathbf{H^c}$ which have large magnitudes should have higher selection priority.**

The above heuristic is a useful first approximation, however, we can improve it by selecting rows uniformly from all the under–represented directions in space. We assume that the rows form clusters that lie in different directions and use an algorithm that detects these clusters, that is, perform unsupervised learning. In particular, we chose a hierarchical clustering algorithm [Hair *et al.*, 1984] that repeatedly decomposes the set of rows into new clusters. Every row belongs to a cluster and every cluster is represented by a seed. At each iteration, first, the row that lies farther from the existing seeds is selected as the seed of a new cluster. Then, the rows that lie closer to the new seed rather than the preexisting ones become members of the new cluster. The decomposition stops when the Euclidean distance between the seeds of different clusters is large enough compared to the scatter within the clusters, or when a maximum number of clusters is reached. We could use any variant of this hierarchical algorithm, a sequential algorithm, or an algorithm based on the principal components of $A^T A$[Cook and Weisberg, 1982].

4 Practical Algorithms and Results

The conditions, the heuristic, and the clustering algorithm presented above were used to implement two practical algorithms that use pixel subsets for recognition. The training set for both algorithms is the SLAM database [Nene *et al.*, 1994]. It consists of 1440 images that correspond to 72 poses of each of 20 objects. Two of the objects are shown in Figures 2(a) and (b).

The recognition algorithms were applied to images of the training set corrupted with Gaussian noise of zero mean; in some images the standard deviation was $\sigma = 10$ and in others $\sigma = 20$. Two of the noisy images are shown in Figures 2(c) and (d). For each noisy image we find the identity of the object. If the object is correctly identified, we estimate its pose. We compute and plot the average of these estimates for 500 test images, for each noise level. In order to compare, we also plot the recognition and pose estimation results for randomly selected subsets of pixels corrupted with noise of the same two levels.

4.1 Window Selection Algorithm

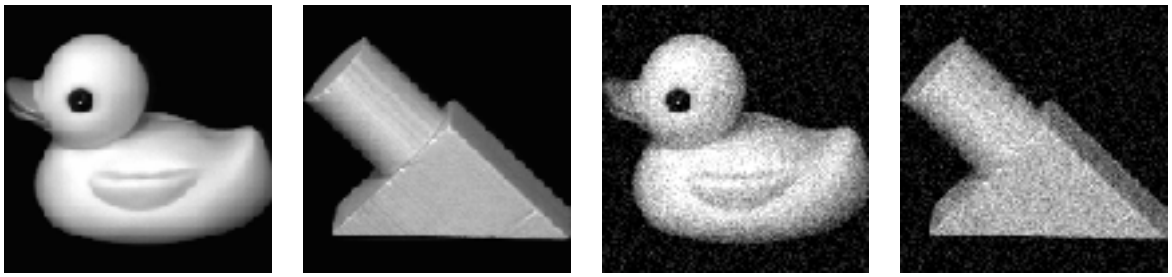
In this algorithm the pixels of the subset are constraint to lie within a window. In addition, the algorithm uses heuristic (1). In the first step of the algorithm we form an image where the intensity is proportional to the corresponding diagonal element of the projection matrix. Then, the window in this image that has the largest sum of intensities in it is used to recognize all the test images.

We show the results for images of size 61×61 in Figures 3(a) and (b). In the figures we vary the size of the window from 10% to 100% of the whole image. The recognition rate obtained with our algorithm is higher than that of the random algorithm. Actually, the rate is 100% for subsets of only 45% of the image pixels. Further, the pose estimated with our algorithm is more accurate than that estimated with the random selection one. For both our algorithm and the random one, pose error is very small for subsets with greater than 60% of the image pixels.

4.2 Pixel Selection Algorithm

In this algorithm the pixels are selected from the entire image. The algorithm has two steps. In the first step we use the hierarchical clustering algorithm we described in the previous section to form different clusters of the rows of A. In the second step, we discard rows from all clusters using heuristic (1).

The algorithm was applied to images downsampled to 21×21 pixels. We plot the results in Figures 3(c) and (d). In the figures the size of the subset ranges from 2% to 40% of the whole image. The recognition rate obtained using our algorithm is higher than that obtained with ran-



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

Figure 2: Images that were used to train and test the two algorithms. In particular, the images in (a) and (b) were used to train the algorithms. The images in (c) and (b) are corrupted versions of those in (a) and (b), with noise of standard deviation $\sigma = 20$, and were used to test the algorithms.

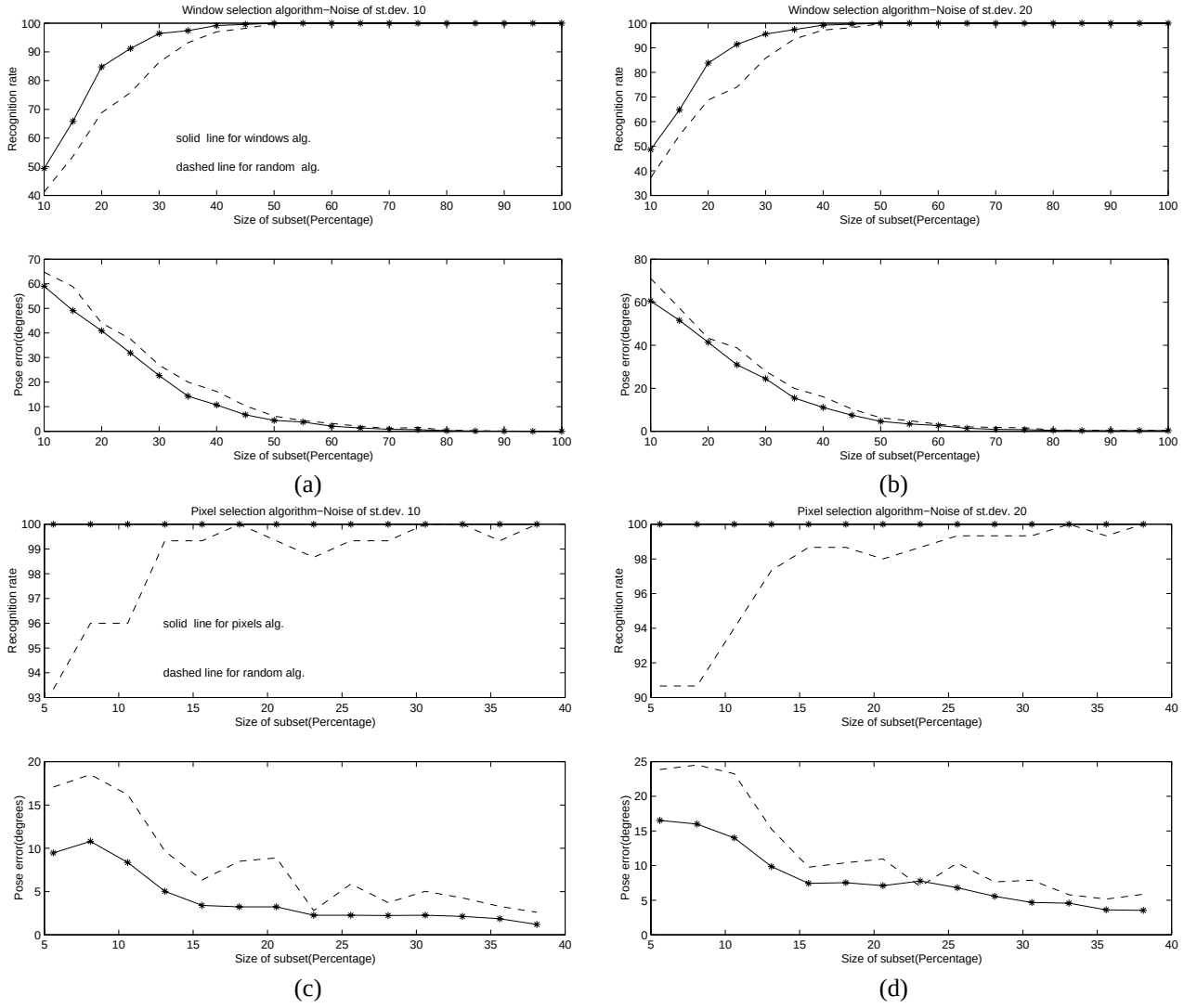


Figure 3: The recognition rate and pose estimation error plotted as a function of the size of the pixel subset. In (a) and (b) we show the results for the window selection algorithm, and in (c) and (d) for the pixel selection algorithm. The plots represent two different levels of noise, namely, $\sigma = 10$ in (a) and (c), and $\sigma = 20$ in (b) and (d). In all experiments, the recognition rate obtained with our algorithms is higher than that obtained with the random selection algorithms. For the pixel selection algorithm it is 100% accurate for subsets consisting of only 6% of the image pixels. Further, the pose errors estimated with both of our algorithms are lower than those estimated with the random ones. In addition, the random algorithms give more incorrectly recognized images, whose erroneous pose is not considered.

domly selected subsets. Actually, our algorithm is 100% accurate for subsets consisting of only 6% of the image pixels. Further, the pose estimated with our algorithm is more accurate than that estimated with random selection.

This algorithm is useful for small images, or when the subsets of pixels are small. If both the image and the subset of pixels are large, the performance of random selection is adequate. This is due to the central limit theorem [Huber, 1981].

Both our algorithms are less sensitive to noise than the corresponding random ones. Hence, the performance of our algorithms compared to the random ones improves for higher levels of noise. Further, both our algorithms have better worst–case performance than the corresponding random ones. This is because the random algorithms can lead to a bad selection of pixels, whereas both our algorithms always lead to a suboptimal subset. Finally, our algorithms are real–time since matrix $\mathbf{H}$, and hence the subsets, can be computed off–line.

5 Conclusion

In this paper, the appearance matching method based on partial data has been enhanced with techniques that are analytically derived. These techniques are applicable in general, they accelerate recognition, and have the potential of recognizing occluded images. They are derived based on low–level sensitivity analysis. The techniques judiciously select a subset of pixels to perform recognition, rather than selecting subsets with ad–hoc arguments or randomly. The validity of the analysis, and its possible applications have been demonstrated experimentally using two practical algorithms.

The sensitivity analysis of the design matrix would be more complete if we included an analysis based on the columns of the design matrix A. The column analysis shows that the design matrix should be orthonormal. Further, the window selection algorithm would be improved if we used a circular rather than a square window. Finally, the pixel selection algorithm could use a better clustering technique.

Acknowledgments

The authors would like to thank Simon Baker of Columbia University for his detailed comments on an early draft that have helped improve the paper.

References

[Box and Draper, 1975] G.E.P Box and N.P. Draper. Robust designs. *Biometrika*, 62:347–352, 1975.

[Brunelli and Messelodi, 1993] R. Brunelli and S. Messelodi. Robust estimation of correlation: An application to computer vision. *IRST Tech. Report*, (9310–15), 1993.

[Cook and Weisberg, 1982] R. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, 1982.

[Ehrenfeld, 1955] S. Ehrenfeld. On the efficiency of experimantal designs. *Annals of Mathematical Statistics*, 26:247–255, 1955.

[Gauss, 1873] C.F. Gauss. *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Werke 4, Section 35, Gottingen, 1873.

[Hair *et al.*, 1984] J.F. Hair, R.E. Anderson, R.L. Tatham, and W.C. Black. *Multivariate Data Analysis with Readings*. 1984.

[Hotelling, 1944] H. Hotelling. Some improvements in weighing and other experimental techniques. *Annals of Mathematical Statistics*, 15:297–306, 1944.

[Huber, 1981] P.J. Huber. *Robust statistics*. Wiley, 1981.

[Krumm, 1996] J. Krumm. Eigenfeatures for planar pose measurement of partially occluded objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 55–60, San Fransisco, California, June 1996.

[Leonardis and Bischof, 1996] A. Leonardis and H. Bischof. Dealing with occlusions in the eigenspace approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 453–458, San Francisco, California, June 1996.

[Moghaddam and Pentland, 1995] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *Proceedings of the 5th International Conference on Computer Vision*, pages 786–793, Cambridge, Massachusetts, June 1995.

[Murase and Nayar, 1995a] H. Murase and S.K. Nayar. Image spotting of 3d objects using parametric eigenspace representation. In *Proceedings of the 9th Scandinavian Conference on Image Analysis*, pages 323–332, Uppsala, Sweden, June 1995.

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

[Nayar *et al.*, 1996] S.K. Nayar, S.A. Nene, and H. Murase. Real–time 100 object recognition system. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2321–2325, Minneapolis, Minnesota, April 1996.

[Nene *et al.*, 1994] S.A. Nene, S.K. Nayar, and H. Murase. Software library for appearance matching (slam). In *ARPA Image Understanding Workshop*, Monterey, California, November 1994.

[Strang, 1980] J. Strang. *Linear algebra and its applications*. Academic Press, 1980.

[Turk and Pentland, 1991] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.