Techical Report: Overview of Appearance Based Methods in Computer Vision

Nicola Duffy

October 6, 1999

Abstract

Appearance based methods turned traditional computer vision approaches to object recognition upside down. Instead of going from image to $3D$ model and matching that model to models of know objects, appearance based methods stay in the $2D$ image domain. Objects are modelled based on how they can appear in images, and this is learned directly from images; it is not necessary to have any explicit knowledge of the shape or reflectance properties of the object.

This technical report presents an overview of appearance based methods in computer vision.

1.1 Appearance Space

In 1991 the question 'What are the fundamental substances of vision?' was asked by Adelson and Berger [AB91]. In answering it, they developed the plenoptic function, P. This function describes everything that can be seen. The plenoptic function returns the intensity level of the reflected light at each point in the picture plane x, y as a function of the wavelength of the light λ, the time, t, and the viewing position, V_x, V_y, V_z:

$$P(x, y, \lambda, t, V_x, V_y, V_z) \tag{1.1}$$

If the plenoptic function of an object is known then every possible image of that object can be constructed. The significance of the plenoptic function is described by Adelson and Berger as follows [AB91]:

> The world is made of three dimensional objects, but these objects do not communicate their properties directly to an observer. Rather, the objects fill the space around them with the pattern of light rays that constitutes the plenoptic function, and the observer takes samples form this function. The plenoptic function serves as the sole communication link between physical objects and their corresponding retinal images. It is the intermediary between the world and the eye.

The problem is in finding the plenoptic function for an object. Exactly how to do this is not evident. What has been the approach of much research is to model the subspace of image space in which the images generated by plenoptic function for an object have their range. Images are considered as high dimensional vectors and the set of all the possible images of an object form a subspace in the space of all possible images. The subspace of an object is called it's appearance space.

Belhumeur and Kriegman [BK98] examined what the appearance space of an object would look like from one viewing position under all possible lighting conditions. They found that if the object has a convex shape and a Lambertian surface reflectance function that the set of n-pixel images produced by all possible lighting conditions is a convex polyhedral cone in $\Re^n$. They show that the the dimension of the cone is equal to the number of distinct surface normals of the viewed object visible from that viewing position and this implies that the cone lies in a very low dimensional linear subspace of image space. This is because the image space has a very high dimensionality, the number of pixels in an image. They also show that the cone can be calculated using just three different images of the object.

The work of Belhumeur and Kriegman has shown that although the appearance space of an object lies in the very high dimensional image space it only actually occupies a much lower dimensional subspace of image space. One way of reducing the dimensionality of the appearance space of an object is to apply the Karhunen-Loéve expansion to the set of possible images of an object (or a representative sample of them). The Karhunen-Loéve expansion has been proven to be a powerful tool in information compression and has been shown to have two very important properties; error-minimisation and entropy-minimising [Wat65]. What the Karhunen-Loéve expansion does is transform the appearance space into a new coordinate system, in which the variance in the images is pushed as much as possible in to the smallest number of dimensions. It is not the only way to reduce the dimensionality of the appearance space but one of the most commonly used.

1.2 The Karhunen-Loéve Expansion

The Karhunen-Loéve expansion, abbreviated the K-L expansion, is a well established method for the compression of the information in a set of continuous functions into fewer and more important variables. When a set of functions is to be expressed as a series in terms of orthogonal and normalised base functions, the K-L expansion minimises the average error induced by taking only a finite number of these functions. In the discrete case the functions are replaced by vectors and the K-L expansion minimises the average error induced by using only a subset of the total set of orthonormal vectors needed to reconstruct a set of vectors. It can also be shown in both the discrete and continuous cases that the K-L expansion minimises the entropy of the average squared coefficients of an expansion. Here average means the average value that the coefficient will take when it reconstructs all the different members of the set. The square of the coefficient is used since it is equivalent to minimising the entropy of the average of the coefficients and makes the proof a little easier. Entropy is the measure of information in the knowledge of the occurrence of an event, or in this case the amount of information in that dimension. As a reminder, Shannon defined entropy as follows [Sha48]:

Definition 1 *Given a set of possible events,* $p_1, p_2, \ldots, p_n$*, where* $\sum_{i=1}^{n} p_i = 1$ *the* **entropy** *is defined as*

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{1.2}$$

Notice that the entropy of a distribution is maximised if all the events are equally probable, the amount of information in the knowledge that a particular event has happened is maximised if all the events were of equal probability. In that case knowing Alternatively the entropy is minimised if all the probability is in one event, that is if for a particular event k, $p_k = 1$ meaning all other events have

a zero probability. In that case the occurrence of event k does not contains any information. The entropy minimisation property of the K-L expansion means that the K-L transform forces the average squared value of the coefficients to be as 'bunched up' as possible. These two properties, error and entropy minimisation, imply that the K-L expansion is the optimal way to compress information. For complete proofs of both these statements see [Wat65].

1.3 Applying the Karhunen-Loéve Expansion to Sets of Images

The K-L expansion was first applied to sets of images by Murakami and Kumar [MK82]. Interpreting images as samples from a stochastic process they used the K-L expansion to extract a set of basis images. The basis vectors, or as in this case basis images, can also be called the principal components and applying the K-L expansion is often referred to as principal components analysis, PCA.

1.4 Calculating the Eigenimages

The first step when applying the K-L expansion to set of images is to calculate the average image of the set and subtracted this average image from each image. The covariance matrix of the images is calculated and then the eigenvectors of the covariance matrix are found. The eigenvalues are equal to the variation in the images that each eigenvector accounts for. Each image can then be exactly reconstructed by a linear combination of the eigenvectors and thereby represented by the coefficients of the linear combination. Using a subset of the eigenvectors, those which have the largest eigenvalues, the best approximate image, in the least squares sense, using that particular number of dimensions can be reconstructed.

Given a set of M images $\vec{\phi}_1, \vec{\phi}_2, \ldots, \vec{\phi}_M$, of size $N = m$ by n, the average image, $\vec{\phi}_{avg}$, is calculated then each image is normalised by subtracting the average image. Subtracting the average image ensures that the eigenvector with the largest eigenvalue corresponds to the dimension in eigenspace with the maximum variance of the images in the correlation sense [MN93]. Let $\vec{\Phi}_i = \vec{\phi}_i - \vec{\phi}_{avg}$. Then a matrix A can be defined as follows:

$$A = [\vec{\Phi}_1 \vec{\Phi}_2 \ldots \vec{\Phi}_M] \tag{1.3}$$

A is a N by M matrix. The covariance matrix, C, is equal to $\frac{1}{M} A A^T$. Even if the images in the set are small this covariance matrix will be very large. The covariance matrix will be of size N by N (remember that N is the number of pixels in an image). As an example consider using a set of images, where the images are sized 64 by 64, which in practice are quite small. The covariance matrix will be of size 4096 by 4096 (16,777,216). Very quickly the computational expense makes calculating the eigenvectors of this matrix infeasible. The rank of this matrix is however bounded by the number of images in the training set. In most situations, there will be many less images than pixels in an image. To avoid the wasted computation of calculating zero eigenvectors Murakami proposed two reduced computation techniques, details of which can be found in [MK82]. The first technique is able to extract only the non zero eigenvectors and the second technique extracts the eigenvectors in order of the value of the eigenvalue. This means it can be halted when the required number of eigenvectors have been extracted.

Another approach to cut to the computation time was proposed by Turk [TP91, Tur91]. The eigenvectors, $\vec{u}_i$, and their corresponding eigenvalues, λ_i, of the covariance matrix are defined as follows:

$$\frac{1}{M} A A^T \vec{u}_i = \lambda_i \vec{u}_i \tag{1.4}$$

Turk noticed that instead of solving the above, the following could be solved:

$$\frac{1}{M}A^T A\vec{v}_i = \mu_i \vec{v}_i \tag{1.5}$$

$A^T A$ is a M by M matrix, where M is the number of images. Then if both sides are pre-multiplied by A this becomes the following:

$$\frac{1}{M}AA^T A\vec{v}_i = \mu_i A\vec{v}_i \tag{1.6}$$

This means that the eigenvector $A\vec{v}_i$ is equal to $\vec{u}_i$.

Calculating the eigenvectors in this way replaces the computation costs, from the order of the number of pixels in the image to the number of images in the training set, often a huge computational saving.

1.5 From Principal Components Analysis to Appearance Space

1.5.1 Face Image Reconstruction

The power of principal components analysis for images was first demonstrated when it was applied to human faces by Sirovich and Kirby [SK87, KS90]. They used PCA analysis for image compression. First a set of eigenimages is learned and then face images, including ones not represented in the training set, were compressed by projecting them into the 'face space' and storing the weights. If k eigenvectors are used, the image, $\vec{I}$, can be projected into that space to give a vector of weights, $\vec{w}$.

$$\vec{w} = [\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k]^T \cdot (\vec{I} - \vec{\phi}_{avg}) \tag{1.7}$$

Using the eigenimages and the weights, face images could then be reconstructed.

$$\vec{I}_r = [\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_k] \cdot \vec{w} + \vec{\phi}_{avg} \tag{1.8}$$

Here the reconstructed image, $\vec{I}_r$, may be an exact copy of $\vec{I}$ if all the eigenvectors are used. If however only a smaller set of those with the largest eigenvalues are used, $\vec{I}_r$ will be the best approximation to $\vec{I}$ for any basis with that number of dimensions. Sirovich and Kirby studied the error effects of using different numbers of the eigenvalues on the reconstructed faces. The faces images used where of size 91×51 and they found that using the first 40 eigenvectors resulted in reconstructed images which humans felt had a reasonable likeness, and this included faces not in the original set. They report a RMS pixel-to-pixel error of approximately 2%, which represents compression ratio of 1:100.

1.5.2 Face Recognition

In 91 Matthew Turk used PCA for face recognition [TP91, Tur91]. The system is initialised in the following manner; a set of face images is acquired, the face images are normalised and aligned, the eigenvectors of this set are calculated and a certain number of those with the highest eigenvalues retained. The weights for each individual are calculated by projecting back onto this space. The weight vectors, $\vec{\Omega}_k$, actually stored for an individual, k, are the average weights calculated for that individual over several images with slight changes in facial expression, pose and lighting. In Turk's experiments 1 to 4 face images for each individual were used.

New images are recognised by projecting in to the face space. The distance to face space is first calculated and a threshold is applied to check to see if the presented image is a face. This distance, ϵ, is calculated by using the distance from the input image, $\vec{I}$ and the image that can be reconstructed of this image

using the face space, $\vec{I}_r$.

$$\epsilon^2 = \| \vec{I} - \vec{I}_r \| \tag{1.9}$$

If this test is passed, that is if ϵ^2 is below a predetermined threshold level, the distance to the nearest known face is calculated. First the vector of the weights of the input image, $\vec{\Omega}_I$, are calculated. Then the k class that minimises the Euclidean distance ϵ_k is found.

$$\epsilon_k = \| \vec{\Omega}_k - \vec{\Omega}_I \| \tag{1.10}$$

If this is above a threshold value the face is classified as an unknown face, otherwise it is classified as the k^{th} individual.

One of the questions that Turk posed was how to quantify the number of eigenvectors needed to give accurate recognition rates. He supposed that if the number of individuals to be recognised was small then a relatively small number of eigenvectors would suffice. As the number of individuals grows then so also the number of eigenvectors. Experiments where done using a set of 16 individuals and an eigenspace with seven eigenvectors.

1.5.3 Generalised Object Recognition

The next step toward appearance space was suggested by Murase and Nayar [MN93, MN95]. They developed a continuous compact representation of an object's appearance which they referred to as the parametric eigenspace. This method goes a step beyond the face compression and recognition methods presented above, which lie in the domain of pattern recognition by building a complete parametrised model of the object. This means that not only the dimensions of an appearance space are learned but also the surface (coefficients) that all possible appearances of an object occupies. In [MN95] they also prove that

an eigenspace representation of an image is optimal in the correlation sense, that is that two distance between two images in the eigenspace corresponds to their similarity under the l_2 norm. The proof of this can be read in [MN95].

A set of images, varying in pose and illumination, was taken for each object to be recognised. The images were normalised by segmenting out the background and reassigning it a zero intensity, and also adjusting the size of the object, so that the objects' largest dimension was always equal. The principal components of all these images were calculated. A hyper-surface in this eigenspace for each object was created using all the images of that particular object and projecting them onto the eigenspace. Using a hyper-space, which was parametrised by pose and illumination, the weights give a selection of discrete points. These discrete points can be interpolated to give a hyper-surface. New objects can be recognised by projecting them onto the eigenspace to get weights and then matching to the hyper-surface that the weights lie on or are closest too. If they are more than a certain distance away from any of the learned hyper-surfaces they are assumed to not be recognisable images. Using a hyper-surface approach is more reliable than using the points as clusters due to the fact that the hyper-surfaces of objects can be intertwined and even intersect each other. The pose of the object can also be determined by it's closest position on the hyper-space of an object.

Experiments done using this method showed that an eigenspace with ten dimensions performed well. With less than four dimension the system was not able to distinguish between the objects but using more than ten did not improve the performance significantly. Again the authors speculated that the number of eigenvectors needed is related to the complexity and total number of objects to be recognised.

A working example of this method is given by Martin and Crowley in the recognition of hand gestures [MC97]. The appearance manifolds for six differ-

ent hand gestures were constructed and the system was able to recognise these gestures.

The work of Murase and Nayar [MN93] was ground breaking for the following reason. Their algorithm was able to automatically learn to recognise objects of interest. Several problems remained; first is a way to automate the selection of the number of eigenvectors to use. Experiments have shown that the more complicated the objects or the greater the number of objects to be differentiated the higher the number of eigenvectors needed, but this hasn't been quantified. A difficulty in the approach is also that of learning new objects. Calculating the eigenspace is very expensive. If a system is to be expected to learn new objects it will either have to re-calculate the eigenspace each time a new object is learned, or just use the now non optimal existing eigenspace. Another limitation is that the object needs to be segmented, which although easy in the case of the controlled environment images dealt with in the paper, not always so easy in more general cases. The method also is not able to recognise objects that have occlusions.

If the basis vectors are used to reconstruct an object that is partly occluded the vectors will also try to reconstruct whatever it is that is actually occluding the true object. This makes the reconstruction of the part of the actual object not as accurate as it could be if the occlusion was ignored. This is exactly what Black and Jepson suggest doing [BJ96]. They reformulate the reconstruction problem as one of robust estimation, by noting that the least-squares image reconstruction of the eigenspace method in the case of occlusion has some problems. Instead of minimising the following, traditional least squares error:

$$E(\vec{c}) = \sum_{j=1}^{n \cdot m} (I_j - \sum_{i=1}^{t} c_i U_{i,j})^2 \tag{1.11}$$

where $\vec{c}$ is the vector of weights calculated by projecting the image, $\vec{I}$, on to the

t eigenvectors, $\vec{U}$, the following is minimised

$$E(\vec{c}) = \sum_{j=1}^{n \cdot m} \rho((I_j - \sum_{i=1}^{t} c_i U_{i,j}), \sigma) \tag{1.12}$$

where σ is a scale parameter and ρ is defined as follows:

$$\rho(x, \sigma) = \frac{x^2}{\sigma^2 + x^2} \tag{1.13}$$

This error function de-emphasises large residual errors. At what size errors are ignored is controlled by σ. Given a reconstructed image $\vec{I}^*$ residuals where

$$|I_j - I_j^*| > \frac{\sigma}{\sqrt{3}} \tag{1.14}$$

are ignored. By grouping pixels that were rejected in the error calculation and repeating the procedure on these pixels this method was successfully able to reconstruct images made up of one third of one learned object and two thirds of another learned object. Using the traditional technique a strange image that was a ghostly mix of the two was constructed. This work certainly offers one approach to deal with occlusion in images. Local appearance space methods offer another way of dealing with occlusion.

1.6 Towards Local Appearance Space

Using whole images to build the appearance space is sensitive to occlusion and works well only if objects in images are of the same size, aligned, and segmented from the background. These criteria pose a huge limitation in applying appearance space methods to real world problems. One solution is to learn the appearance of local areas in images instead of appearances of whole images.

1.6.1 Eigenfeatures

In 1996 John Krumm developed a system for measuring the pose of objects on a plane, such as a conveyor belt [Kru96]. The system is an extension of the parametric eigenspace method of Murase and Nayar discussed above, but does not require segmentation of the object in the image, because local patches are used. Objects are viewed from above and rotated while images are captured every 2 or 4 degrees. A feature detector is then run over the image, examining image patches of size 15 by 15 pixels. Instead of applying PCA to the whole images, the image patches which scored highly with the feature detector, are used to create a low dimensional appearance space. Every pixel in the image is used as the centre of a patch and back projecting the patch into the appearance space. The pixel values are replaced with the weights vectors. The image is then scanned and features in recognisable configurations are sought. Using the patches enables this system to be tolerant to partial occlusion.

1.6.2 Eigen-Windows

Another approach in the same spirit is given by Ohba and Ikeuchi [OI96]. They refer to it as the Eigen-Window method. They examine all the possible windows for an image and as in the method of Krumm they select windows that pass a feature detector but then they also remove similar windows. Removing similar windows not only makes the training set smaller and thereby reducing memory and computational costs but it also makes the matching process more robust since the set only contains unique windows.

1.6.3 Local Appearance Space

The method proposed Colin de Verdiére and Crowley [dVC98b, dVC98a] uses all the possible small neighbourhood windows (or imagettes) to construct an appearance manifold. If the imagettes are $m \times m$ then the dimensionality of the local appearance space will be m^2. An image is projected into this space by taking all of the possible m by m imagettes. Since there will be a large overlap in neighbouring imagettes they will be highly correlated and therefore lie very close to each other in local appearance space. This means that the projection of an image into local appearance space gives a discrete sampling of the surface of appearance of an object. Different views of an object give a family of surfaces.

An interesting observation made by Colin de Verdiére was that the set of images used to create the local appearance space had surprisingly little effect on the eigenvectors resulting in PCA. In fact the eigenvectors look very similar to Gaussian derivate filters, which would suggest that a set of universal filters, approximately optimal for any image set, exists, see figure 1.1, which shows filters that were actually calculated in the experiments. Although local appearance space methods were motivated by the desire to be tolerant to occlusion and non-segmentation, it resulted in a suggestion for the solution to the problem of eigenspace updating, since the optimal eigenvectors do not change significantly.

Training

Training was done by using images from the Columbia Image Database [NNM96], which contains 7,200 images of 100 objects. There are 72 images of each object taken against a black background every 5 degrees, under constant illumination. The training images were chosen by selecting images taken set every 20 degrees, leaving the remaining images available for testing. Taking training images every

20 degrees may not be optimal for some of the objects, those which change rapidly under rotation, but this level of sampling proved to be reasonable over the whole data set.

The imagette size used was 9×9, giving a maximum dimensional appearance space of 81. The size of 9×9 was chosen since it was the smallest size for which the local spatial frequency is not perturbed by the sinc function which is inherent in neighbourhood selection [dVC98b].

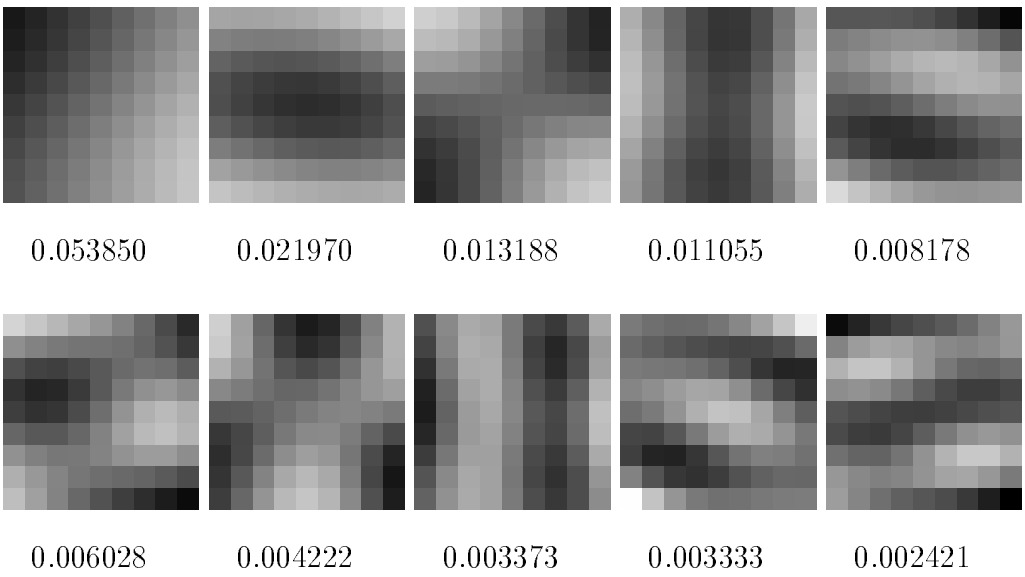

Figure 1.1: Eigenvectors and their corresponding Eigenvalues [dVC98b]. Note that these images have float and negative pixel values and have been self-normalised to the range $[0, 255]$ to be printed.

Recognition

Recognition is achieved by taking a window from an unknown image, projecting it into the local appearance space and associating it with the object whose surface it lies closest to. As confidence factor can be calculated by combining the distance to the chosen surface and the number of nearby surfaces. Using just one 9×9

window good recognition rates where achieved but excellent results were achieved using 2 windows and also requiring that the spatial correspondence between the two windows corresponded to the learned spatial correspondence.

Extending Local Appearance Space Methods to Colour

The images in the Columbia Image database are in colour. In order to include the colour information of the image Colin de Verdiére used imagettes that were 3 dimensional instead of 2 dimensional. That is instead of imagettes of m by m they were m by m by 3. In terms of the calculations these imagettes are treated like m by 3m. Extending this approach to use colour in this way gave even more robust recognition results than gray-scale local appearance based matching.

1.7 Conclusion

Appearance based methods represent a fundamental change in computer vision approaches. The learning phase is computationally expensive and updating the filters as new objects are learned is expensive. The work of Colin de Verdiére and Crowley has suggested this if local appearance spaces are used, then a set of universally optimal filters could make it unnecessary to calculate the filters instead learning would only consist of constructing the hyper-surface of each object. These techniques are still quite new and a more through exploration to the full extent of their possibilities and limitations is needed. In the next chapter several methods to detect the burn marks in the filter casings using local appearance space methods are presented and the experimental results given.

Bibliography

[AB91] Edward H. Adelson and James R. Bergen. The plentoptic function and the elements of early vision. *Computational Models of Visual Processing*, pages 3–20, 1991. http://www-bcs.mit.edu/people/adelson/publications/abstracts/elements91.html.

[BJ96] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV'96 Fourth European Conference on Computer Vision*, pages 329–342. Lecture Notes in Computer Science, Springer Verlag, 1996.

[BK98] Peter N. Belhumeur and David J. Kriegman. What is the set of images of an object under all possible illumination conditions. *International Journal Of Computer Vision*, 28(3):245–260, March 1998.

[dVC98a] Vincent Colin de Verdiére and James L. Crowley. Local appearance space for recognition of navigation landmarks. In *The Sixth Symposium for Intelligent Robotics Systems*, July 1998.

[dVC98b] Vincent Colin de Verdiére and James L. Crowley. Visual recognition using local appearance. In *The Fifth European Conference on Computer Vision*. Lecture Notes in Computer Science, Springer Verlag, June 1998.

[EBW89] S. Edelman, H. Bulthoff, and D. Weinshall. Stimulus familiarity determines recognition strategy for novel 3d objects. A.I. Memo No. 1138, A.I. Lab, MIT, 1989.

[Kru96] John Krumm. Eigenfeatures for planar pose measurement of partially occuled objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 55–60, June 1996.

[KS90] M. Kirby and I. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, January 1990.

[MC97] Jerome Martin and James L. Crowley. An appearance-based approach to gesture-recognition. In Alberto Del Bimbo, editor, *International Conference on Image Analysis and Processing*, Florence, Italy, September 1997. Lecture Notes in Computer Science, Springer Verlag.

[MK82] Hiroyasu Murakami and B.V.K. Vijaya Kumar. Efficient calculation of primary images from a set of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(5):511–515, September 1982.

[MN93] Hiroshi Murase and Shree K. Nayar. Learning and recognition of 3d objects from appearance. In *Proceedings of the IEEE Workshop on Qualitative Vision*, pages 39–50, New York, June 1993.

[MN95] Hiroshi Murase and Shree K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[NNM96] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library. Technical report, Columbia University, New York, February 1996.

[OI96] Kohtaro Ohba and Katsushi Ikeuchi. Recognition of the multi specularity objects using the eigen-window. Technical Report CMU-CS-96-105, Carnegie Mellon University School of Computer Science, Pittsburgh, Pensylvania, February 1996.

[Sha48] Claude Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, July and Octber 1948. http://www.math.washington.edu/ hillman/entropy.html.

[SK87] I. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc Am.*, 4(3):519–524, March 1987.

[TP89] M. Tarr and S. Pinker. Mental rotation and orientation-dependance in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuro Science*, 3(1):71–86, 1991.

[Tur91] Matthew Turk. *Interactive-Time Vision: Face Recognition as a Visual Behavior*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1991. Advisor: Alex P. Pentland.

[Wat65] Satosi Watanabe. Karhunen-loeve expansion and factor analysis theoretical remarks and applications. In *Trans. Fourth Prague Conference Information Theory, Statistical Decision Functions, and Randon Processes*, pages 635–660, Prague, 1965.