MVA2000 IAPR Workshop on Machine Vision Applications, Nov. 28–30, 2000, The University of Tokyo, Japan

8—3
Yet Another Appearance-Based Method for Pose Estimation Based on a Linear Model

Takayuki Okatani and Koichiro Deguchi
Graduate School of Information Sciences, Tohoku University
{okatani, kodeg}@fractal.is.tohoku.ac.jp

Abstract

This paper explores the possibility of a linear model as a solution to the problem of appearance-based pose estimation. The parametric eigenspace method (or its extensions that are based on correlation between images) has been widely used and yields successful results for the pose estimation problem. On the other hand, the method has some problems. One is that large computational cost and storage space are required. Another is that small changes in appearance can be discarded even if it is related to changes of parameters to be estimated. Based on these, another appearance-based method for estimating pose of an object using a linear model is examined. Experimental results are not superior to the eigenspace method in terms of estimation accuracy. However, it has several advantages to the parametric eigenspace method in terms of storage space and computational cost, and has several features that can be advantages to the eigenspace method.

1 Introduction

The parametric eigenspace method proposed by Murase and Nayar has led to a wide range of successful applications [1]. Although it was originally proposed for the problem of estimating the pose of an object from its appearance, it has been widely used for more general purposes to estimate physical parameters underlying in appearance of an object or a scene. (In [1], the problem of object recognition as well as pose estimation is treated; the process of object recognition is followed by the process of pose estimation. The focus of this paper is on the part of pose estimation after object recognition is completed.) At least formally, the problem can be written as follows: when the image $\mathbf{x}$ and some physical parameters $\mathbf{p}$ have a certain unknown relation, $\mathbf{x} = \mathbf{x}(\mathbf{p})$, estimate $\mathbf{p}$ for a given $\mathbf{x}$. The relationship $\mathbf{x} = \mathbf{x}(\mathbf{p})$, which is usually nonlinear, is learned using a training set of N samples, $\{(\mathbf{x}_i, \mathbf{p}_i) \mid i = 1, \ldots, N\}$.

The parametric eigenspace method is a correlation-based method. Its name comes from that images are represented in their eigenspace for data compression, and the data compression by the eigenspace itself has nothing to do with improvement of estimation accuracy. The process of estimating parameters is summarized as follows: when an image $\mathbf{x}$ is given, the estimation for $\mathbf{p}$ is determined from the samples, $\{(\mathbf{x}_i, \mathbf{p}_i) \mid i = 1, \ldots, N\}$, in a way that the nearest samples are first chosen and then the parameter is interpolated using the parameter values of the nearest samples ($\mathbf{p}_j$) based on the difference in appearances. The reason why such appearance-based method works well is because appearances of an object for two close poses are usually close to each other, and the changes in appearance due to pose changes is continuous.

Although it will not contribute in a positive way of improving estimation accuracy of parameters, the data compression using eigenspace is crucial for real implementations, since storing samples and computing differences between samples as raw images are exhaustive. Due to this necessity of the data compression, the parametric eigenspace method becomes feasible, and at the same time, it has several problems.

One is discard of image information that is possibly valuable for estimating the pose parameters. The eigenspace is a subspace constructed based only on appearances of the samples and it has no relation to the parameter values of the samples. A sample is given in the form $(\mathbf{x}_i, \mathbf{p}_i)$, and therefore each $\mathbf{x}_i$ is "labeled" by the parameter $\mathbf{p}_i$. However, the eigenspace is constructed without using this. It is possible that the parameters do change and nevertheless corresponding appearance change is small. In such a case, it is important to extract that small appearance change in order to estimate the parameters correctly. However, such small appearance change can be neglected in the eigenspace if appearance change due to other parameter change is dominant in the samples. This is similar to the discussion in the problem of the object recognition that the feature space should be chosen so that the distance between classes becomes maximum [2].

The other is the problem of large computational cost and storage space. Although we can reduce them by representing samples in the eigenspace, it can be still large when plural parameters are to be learned and estimated. For example, when 3 parameters are to be estimated and 100 poses for each parameter have to be learned, the number of total samples is 100^3. Due to data compression by eigenspace, it becomes feasible to store such a

large number of samples in memory and to access them. Nevertheless, they are still large. (Even if the dimension of the eigenspace is reduced to 10, $100^3 \times 10$ floating point numbers have to be stored in memory.)

Taking these problems into account, we present another way of estimating object pose from its appearance. The present method is based on a linear model of the relation between appearance and the pose parameters. Although it is fairly simple and thus it seems trivial, we think that the associated problems are interesting and suggestive.

2 Possibility of linear models

In this section, yet another appearance-based method is presented. The parameters $\mathbf{p} = [p_1, \ldots, p_K]^\top$ are estimated based on a linear model:

$$p_j = \mathbf{w}_j^\top \mathbf{x} \quad (j = 1, \ldots, K). \tag{1}$$

Taking a training set of N samples, $\{(\mathbf{x}_i, \mathbf{p}_i) \mid i = 1, \ldots, N\}$, we determine $\mathbf{w}_j$ so that the above model holds for the samples. Clearly $\mathbf{w}_j$ is determined as a solution to the following equation:

$$[p_{1j}, p_{2j}, \ldots, p_{nj}]^\top = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \mathbf{w}_j, \tag{2}$$

where p_{ij} denotes the jth element of the vector $\mathbf{p}_i$, i.e., the jth parameter of the ith sample.

The number of image pixels, M, is always large (10,000 – 100,000). The number of the samples, N, is generally much smaller than M. (It is impossible to prepare as many sample images as M!) Thus, $M \gg N$. Therefore, the number of unknowns is larger than that of equations, and an infinite number of solutions can exist. In many possible solutions, we choose $\mathbf{w}_j$ so that $|\mathbf{w}_j|$ is minimum. Let X be $[\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$. Such a solution is represented as $\mathbf{w}_j = X\mathbf{d}_j$ using $\mathbf{d}_j$ that satisfies

$$[p_{1j}, p_{2j}, \cdots, p_{Nj}]^T = X^\top X \mathbf{d}_j. \tag{3}$$

If the square matrix $\mathbf{X}^\top \mathbf{X}$ is nonsingular, $\mathbf{w}_j$ is written by

$$\mathbf{w}_j = X(X^\top X)^{-1}[p_{1j}, p_{2j}, \cdots, p_{3j}]^T. \tag{4}$$

Then, using this $\mathbf{w}_j$, the parameter for given $\mathbf{x}$ is estimated by Eq.(1). The determination of $\mathbf{w}_j$ and the estimation of p_j are conducted for each j.

Whether $X^\top X$ is nonsingular or not is entirely dependent on the set of the sample images. Since the correlation between the sample images is usually high, the eigenvalues of $X^\top X$ are biased; only a small number of eigenvalues are dominant. This is a basis for the eigenspace method. However, we have confirmed through experiments that even the smallest eigenvalue does not become zero in usual situations. Of course, if different samples have the same appearance, then 0 eigenvalue occurs and $X^\top X$ becomes singular. In this case, however, estimating parameters from appearance is impossible in principle. Furthermore, the bias of the eigenvalues is almost due to the resemblance of neighboring sample images. Although neighboring sample images are close with each other, corresponding parameters are also close with each other. Therefore, $\mathbf{w}_j$ satisfying Eq.(2) usually exists. (Only by an experiment using real images, this can be examined.)

Table 1: Comparison between the parametric eigenspace method and the present method. L is the dimension of the eigenspace. *) The case for the nearest neighbor search algorithm.

	Eigenspace Method	Linear Model
Storage Space	$NL + ML$	MK
Comput. Cost	$NL + ML^*$	MK

The present method has an advantage to the eigenspace method in terms of storage space and computational cost. Table 1 shows the comparison. (L denotes the dimension of the eigenspace.) In the eigenspace method, each sample is stored as a look-up table (called the appearance manifold in [1]) in the eigenspace, and therefore the storage space required is usually larger. The storage space required for samples is order of NL. Along with that for the base vectors of the eigenspace, the total space required is $NL + ML$. Although this space is much smaller than storing samples as raw images, the storage space required is still not small, since a large number of samples are usually necessary. (Murase et al. aimed at reduction of the storage space using a spline surface fit.) Besides, computational cost is not small. If the nearest neighbor search algorithm is employed, it amounts to $NL + ML$. Several methods for reducing this cost have been proposed, one of which is to use a neural network learning the nonlinear appearance manifold. It is not clear, however, that such a usage of neural networks is definitely superior to the linear model we have described. On the other hand, the present linear method does not require such a look-up table; storage space as well as computational cost is basically M for each parameter.

3 Experimental results

We show here experiments of estimating pose of an object using three objects shown in Fig.3. For each of the objects, 1760 images of 160×120 pixels under different poses were taken. The pose of the objects is represented by the polar and azimuthal angles, `q` and `f`. To acquire samples, a turntable is used for changing `f` and a robot hand is used for changing `q`. A CCD camera is mounted on the robot hand and takes images. The samples are obtained by changing `f` from 0 to 330 degrees

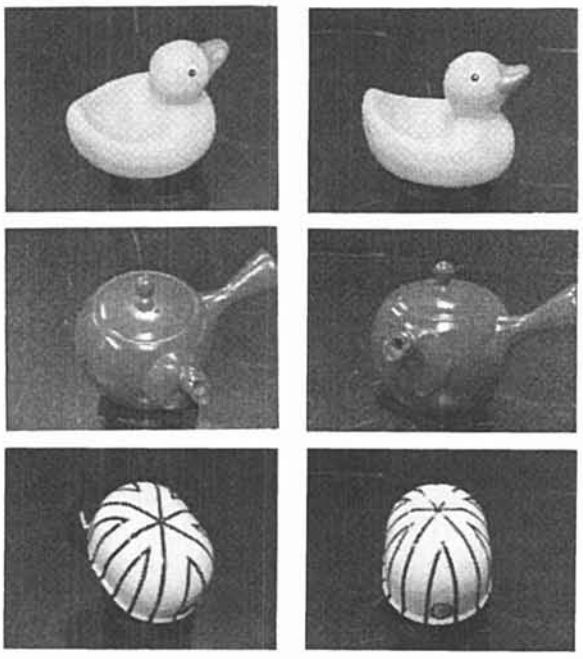

Figure 1: Three objects used for the experiments: DUCK, POT, and CAP. The first and the last images in the samples are shown for each object.

by 3 degrees, and by changing $\mathtt{q}$ from 32 to 80 degrees by 3 degrees. For each object, the number of samples is $110 \times 16 = 1760$.

Since appearance for $\mathtt{f} = 0$ degree and appearance for $\mathtt{f} = 360$ degrees are the same, using $\mathtt{f}$ directly for learning and estimation leads to a problem. Therefore, we employ a redundant expression of the parameters:

$$\mathbf{p} = [p_1, p_2, p_3]^\top \equiv [\mathtt{q}, \cos \mathtt{f}, \sin \mathtt{f}]^\top. \tag{5}$$

Thus, three parameters are to be estimated and are associated with $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$, respectively.

We divided the obtained samples into two sets. One is a training set used for learning, i.e., for determining $\mathbf{w}$, and the other is a test set used for testing the generalization ability. The training set is chosen so that $\mathtt{f}$ and $\mathtt{q}$ varies by 9 degrees. Thus, the number of the samples for learning is $37 \times 6 = 222$. For this learning set, $\mathbf{w}_j$ are determined by Eq.(4) and they are shown in Fig.2.

Figure 3 shows the errors of the pose estimation for all samples including the training set. They show the errors of the angle in degree between the estimated pose and the true pose. It can be seen that the pose is estimated with reasonable accuracy for overall samples. However, the results are not superior to the eigenspace method in terms of estimation accuracy. It is expected that the parametric eigenspace method should yield a bit more accurate results.

4 Discussion on overtraining

The parametric eigenspace method has generalization ability. It is able to estimate parameters with reasonable accuracy for a novel image that is different from any image in a training set. This might be due to the nature of the relation between the pose of an object and its appearance; the correlation between the images associated with near pose parameters is high and continuously changes with the change of parameters.

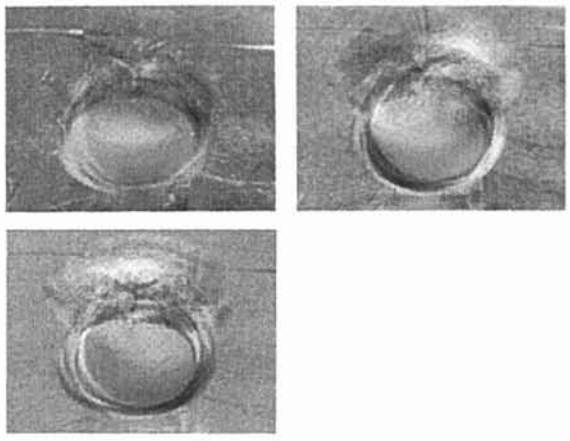

Figure 2: Coefficient vectors for the DUCK sequence that corresponds to $\mathtt{q}$, $\sin \mathtt{f}$, and $\cos \mathtt{f}$, respectively.

As shown in the last section, the linear method has generalization ability, although it is a bit inferior to that of the eigenspace method. In the case where the size of a training set is smaller than the dimensionality of the feature space, it is known that a phenomenon called overtraining occurs. In our problem, the size of training set (e.g., 222) is much smaller than the number of image pixels (e.g., 160×120). It seems that the same problem should occur.

Several studies on overtraining have been done and some results have been obtained so far. One of them which seems to be closely related to our problem is one that assumes the following model:

$$p_j = \mathbf{w}_j^\top \mathbf{x} + n, \tag{6}$$

where n is noise. The parameter p_j is assumed to be derived according to this model. To be estimated is $\mathbf{w}_j$ but its true value cannot be obtained. Then, its estimation, $\hat{\mathbf{w}}_j$, is determined in the same way as Eq.(4) from a training set $\{(\mathbf{x}_i, \mathbf{p})_i\}$. It is shown that if the dimensionality of $\mathbf{x}$ is larger than the size of the training set, then $\hat{\mathbf{w}}_j$ is affected by the noise n and as a result, the generalization error increases with the size of the training set. In our problem, however, the pose parameter p_j can be known precisely and its observation can be made without noise in the step of acquiring the training set. Of course, the relation between the image $\mathbf{x}$ and the pose parameters $\mathbf{p}$ is nonlinear and cannot be completely represented by a linear model. However, it does not seem true that the noise in Eq.(6) accounts for the difference between the linear model and the nonlinear relation. Therefore, it does not seem that the above result is applicable to our problem without modifications. It seems that the nature

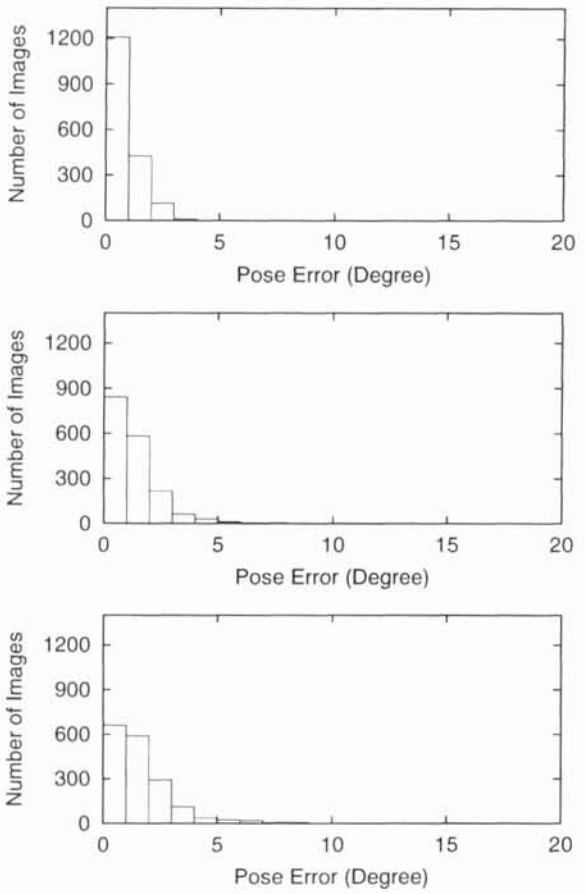

Figure 3: Pose error for DUCK, POT, and CAP (upper to lower row).

of the natural images that the appearance continuously changes with changes in object pose plays an important role. This needs to be explored.

In order to examine the generalization ability of the linear model, we tested the estimation for noisy images. For the DUCK image sequence, using $\mathbf{w}_j$ that is determined for the training set of images without noise (the same images as the results shown in Fig.3), pose estimation is conducted for a test set of images to which uniform noise is added. Uniform noise is added to each pixel's brightness of the image. The results are shown in Fig.4. The noise level is $\pm 5\%$ and $\pm 10\%$ of the maximal image brightness. It can be seen that despite of considerably high noise level, it does not severely affect the results .

5 Conclusion

The possibility of a linear model as a solution to the problem of pose estimation from appearance of objects is examined. The results are inferior to the results that are expected to be obtained by the eigenspace method in terms of estimation accuracy. On the other hand, the linear model has an advantage that is clearly recognizable.

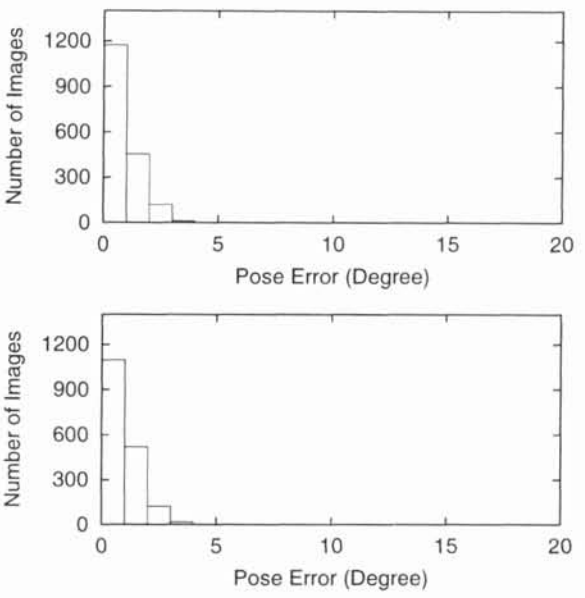

Figure 4: Pose estimation error for noisy images. Added noise is 5% (upper) and 10% (lower).

It is that both storage space and computational cost required are less than the eigenspace method. Furthermore, the linear model uses the pose parameters in the training process and can make full use of image information that the eigenspace method might discard through the data compression.

Several cases that might be a problem for the eigenspace method can be listed. One example is the case where the target object has only slightly different appearances for more than two poses. For example, there can be an object such that the front appearance resembles the backside one. If appearance viewed from any other angle is much different from the front and the backside one, it is possible that the small difference in appearance between the front and backside is not sufficiently represented in the eigenspace. This might be a problem for accurate estimation. Another example is the case where there are two parameters and the appearance changes due to one parameter are much larger than those due to the other parameter. The parameter associated with the smaller appearance changes is not fully represented in the eigenspace. This might be also a problem. These need to be explored furthermore.

References

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

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Patt. Anal. Machine Intel.*, 19(7):711–720, 1997.