Czech Pattern Recognition Workshop 2000, Tomáš Svoboda (Ed.)
Peršlák, Czech Republic, February 2–4, 2000
Czech Pattern Recognition Society

Pose Estimation Using Parametric Stereo Eigenspaces

Thomas Melzer and Michael Reiter
Vienna University of Technology
Pattern Recognition and Image Processing Group
Favoritenstr.9/1832, A-1040 Vienna
melzer@prip.tuwien.ac.at

Abstract *We propose new approach to constructing object models in parametric eigenspace based on stereo pairs. For each object pose, two images are taken from different viewing positions separated by a fixed baseline, and concatenated before they are projected into eigenspace. Experimental results indicate that the stereo approach is superior to its conventional counterpart in terms of accuracy of the estimated pose parameters.*

1 Introduction

In the object recognition community, eigenspace methods have been a very active research topic during the last few years and have led to a variety of successful applications, including face recognition [7], illumination planning [4] and visual inspection [6]. The key idea is not to build an explicit object model, but rather to automatically construct a low dimensional object representation from a set of raw brightness images. Since similar views of the same object tend to be highly correlated, these views can be efficiently compressed using Principal Component Analysis (PCA, [1]). After the covariance matrix of the different object views has been computed, each view can be represented as a linear combination of the eigenvectors corresponding to the largest eigenvalues) of the covariance matrix (typically, 10-20 eigenvectors are sufficient). The coefficients of this linear combination (which are normally obtained be projecting the input image onto the eigenvectors) are the *eigenspace representation* of the object view. Compression techniques that exploit the correlation between signals (e.g., images) by transforming these signals into a low dimensional space, which is spanned by the directions carrying the most information, are usually referred to as *subspace methods*; eigenspaces are just a special case of this more general paradigm that uses linear PCA in order to determine the "most informative" directions.

In contrast to explicit 3D (CAD) object models, the eigenspace representation captures not only geometrical (pose, orientation), but also photometric (e.g., reflectance, texture) properties of the object (*appearance based object representation*) [5]. If these appearance properties are described quantitatively by a set of continuous parameter values, the eigenspace projections of different object views can be thought of as lying on a multidimensional parametric manifold (*parametric eigenspace representation*). An example of such a parametric manifold can be seen in Fig. 1. Eigenspace representations of intermediate views not contained in the training set can be obtained by interpolation, e.g., by using bicubic splines [5] or a Radial Basis Function neural network [3].

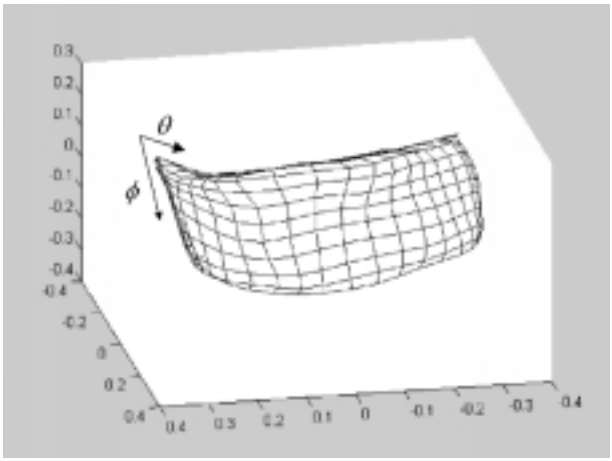


Figure 1: Bivariate parametric manifold describing the Garfield toy figure under various pan (θ) and tilt (ϕ) angles. For visualization purposes, only the projections onto the first three eigenvectors are shown. See section 3 for details.

This work proposes a new approach to constructing object models in parametric eigenspace. For each object pose, two images (a so called stereo pair) are taken from different viewing positions separated by a fixed baseline. As shown in section 3, the use of stereo pairs leads to increased performance as compared to the conventional mono approach in terms of accuracy of the estimated pose parameters.

The rest of this paper is organized as follows. In section 2, we introduce the concept of parametric eigenspaces in general and its extension to parametric stereo eigenspaces. Experimental results, which compare the stereo with the conventional (i.e., mono) approach w.r.t. parameter estimation accuracy, are given in section 3, followed by conclusions in section 4.

2 Parametric Eigenspaces

Principal Components Analysis (PCA) [2, 1] is a well known technique in computer science and statistics for linear feature extraction and dimensionality reduction. The key idea is, given a set of observations drawn from a multivariate distribution **P**, to identify the most informative directions (i.e., those with the largest variance) of this distribution. These directions are called the principal components of the distribution and can be found using KLT (Karhunen-Loéve Transform, see section 2.1 below). The actual feature extraction step for a given sample vector $\mathbf{x}$ is performed by projecting $\mathbf{x}$ onto the principal components, resulting in a new feature vector $\mathbf{y}$.

As explained in more detail below, the principal components of a distribution **P** are given by the the eigenvectors of its covariance matrix. Thus, PCA has three important properties:

- The new features (i.e., the components y_i of $\mathbf{y}$) are "synthetic" features derived from the original data. Thus, in general, they will not correspond to observable physical quantities.
- Since the eigenvectors diagonalize the covariance matrix, the new features y_i are uncorrelated (independent, if **P** is a normal distribution).
- A new feature y_i is obtained as projection of $\mathbf{x}$ onto the eigenvector (principal component) $\mathbf{e}_i$. The relative importance of the feature y_i is given by the standard deviation along the eigenvector $\mathbf{e}_i$, which is in turn given by the corresponding eigenvalue v_i. Thus, dimensionality reduction can be achieved by projecting a data vector $\mathbf{x}$ only onto the eigenvectors with the largest corresponding eigenvalues.

2.1 Eigenspace model

Given N observations[1] $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \Re^n$ an estimate of the covariance matrix Σ of the underlying distribution is given by

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T, \qquad (1)$$

where $\mathbf{m} \in \Re^n, \mathbf{m} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ is the estimated mean vector.

The eigenspace model $\Omega = \langle N, \mathbf{m}, E, V \rangle$ comprises the mean vector $\mathbf{m}$, a set of eigenvectors $E = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}, \mathbf{e}_i \in \Re^n$ of $\hat{\Sigma}$ and corresponding eigenvalues $V = \{v_1 \ldots v_k\}, v_i \in \Re$, that can be obtained by eigendecomposition (or singular value decomposition)

$$\Sigma = \Phi \Lambda \Phi^T, \qquad (2)$$

where $\Phi \in \Re^{n,n}$ contains in its columns the eigenvectors $\mathbf{e}_1 \ldots \mathbf{e}_n$ of Σ and Λ is a diagonal matrix with the associated eigenvalues. Φ is the basis of the *Karhunen-Loéve Transform* (KLT). The KLT rotates the image vectors into a coordinate system, in which the image vector components are decorrelated. The transformed image vector $\mathbf{y}$ is given by $\mathbf{y} = \Phi^T (\mathbf{x} - \mathbf{m})$. The original image vector $\mathbf{x}$ can be reconstructed using the inverse transform $\mathbf{x} = \Phi \mathbf{y} + \mathbf{m}$.

When the number of observations is small compared to the dimensionality of image vectors (i.e. $N < n$) the rank of $\hat{\Sigma}$ is at most $N - 1$ and estimating the complete set of eigenvectors of Σ is not possible. However, in the eigenspace model (which is based on *Principal Component Analysis*) only the $k < N - 1$ largest[2] eigenvectors are retained. These principal eigenvectors can be estimated (using $\hat{\Sigma}$) even when N is much smaller than n.

The principal component feature vector $\tilde{\mathbf{y}} \in \Re^k$ is given by the first k principal components of $\mathbf{y}$ which are the projections onto the principal eigenvectors ($\tilde{\mathbf{y}} = \Phi_k^t (\mathbf{x} - \mathbf{m})$, where we denote the submatrix of Φ containing the first k eigenvectors by $\Phi_k \in \Re^{n,k} = \langle \mathbf{e_1} \ldots \mathbf{e_k} \rangle$). This truncated representation corresponds to an orthogonal projection of $\mathbf{x}$ onto the subspace spanned by $\mathbf{e}_1, \ldots, \mathbf{e}_k$.

In general the original image vector $\mathbf{x}$ cannot be reconstructed completely from $\tilde{\mathbf{y}}$ but only approximated using the inverse transform $\tilde{x} = \Phi \tilde{\mathbf{y}} + \mathbf{m}$. The residual reconstruction error $\epsilon^2(\mathbf{x})$ is equivalent to the Euclidean distance between $\mathbf{x}$ and its projection $\tilde{\mathbf{x}}$ and is to the sum of the squared components of $\mathbf{y}$ that have been omitted in $\tilde{\mathbf{y}}$:

$$\epsilon^2(\mathbf{x}) = \|\tilde{\mathbf{x}} - \mathbf{x}\|^2 = \sum_{i=k+1}^{n} \mathbf{y}^2 = \|\tilde{\mathbf{x}}\|^2 - \sum_{i=1}^{M} \mathbf{y}^2 \qquad (3)$$

The PCA is optimal in the sense that the expected residual reconstruction error for the set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is minimized. It is given by $E(\epsilon^2(\mathbf{x})) = \sum_{i=k+1}^{n} v_i$. Hence, the minimum error is obtained by discarding all but the k largest eigenvectors. The number of eigenvectors k that are used in the eigenspace model Ω can be chosen according to a specified fraction of energy in the eigenvalue spectrum that has to be retained.

2.2 Parametric Stereo Eigenspaces

By explicitly taking into account various parameters that govern an object's appearance (for instance, the viewing angle or position w.r.t. camera system) when building its eigenspace model, these parameters can be retrieved later together with the object's identity in eigenspace. As demonstrated by Nayar et al. [6], such *parametric eigenspace models* can even be used for simple, image driven effector control.

Our hypothesis is that the accuracy of the estimated pose parameters can be increased by providing additional visual information that helps to discriminate between similar poses. Thus, for each object pose, two images (also called a stereo pair) are taken from different viewing positions separated by a fixed baseline. These two views are then concatenated to form a single stereo image vector. Refer to Fig. 4 for an example of a stereo pair.

[1] An $x \times y$ image $\mathbf{X} \in \Re^{x,y}$ can be written as an image vector $\mathbf{x} \in \Re^{n=xy}$ by lexicographic ordering of the image pixels

[2] by "largest" eigenvectors we denote the eigenvectors with the largest corresponding eigenvalues

Note that, although stereo vectors have twice the size of mono image vectors , the resulting eigenspace representations (the eigenspace projections of the reference views) are of size $|R|nf$ in both approaches, whereby $|R|$ is the number of reference vectors, n is the eigenspace dimension and f is the size of a floating point number. Thus, after the image vectors have been transformed into eigenspace, there is no difference in memory or run-time requirements between both approaches; the only additional run-time overhead incurred by the stereo approach is due to normalization and projection of the input image pairs.

3 Experiments

For our experiments, we used toy figures of Garfield, Kemal and Fozzy bear (see Fig. 2). The camera was mounted rigidly on the gripper of an A465 robot. Training and test images were generated by moving the camera center along the surface of a sphere centered around the actual figure with the optical center pointing towards the figure. We used the pan (θ) and tilt angle (ϕ) of the camera w.r.t. the spherical object coordinate system as pose parameters. Stereo pairs at pose (θ_i, ϕ_i) were taken by shifting the camera 15mm along its positive and negative X-axis, starting from pose (θ_i, ϕ_i), respectively, while keeping the orientation of the camera fixed.

Figure 2: The three toy figures of Garfield, Kemal and Fozzy Bear used in the experiments.

This setup differs from the more conventional approach, in which the camera remains stationary, while the object is moved (typically, on a turntable). By making use of an active, effector mounted sensor, however, we can obtain object views under up to three degrees of controlled rotational freedom and thus build a more general object model. The setup used in the experiments is illustrated in Fig. 3.

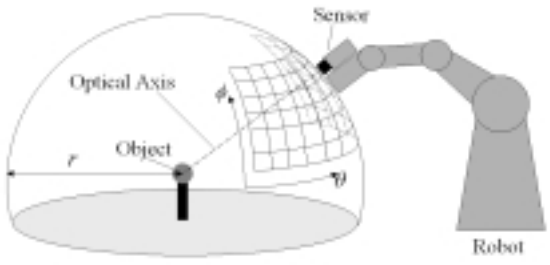


Figure 3: The setup used in the experiments. Images of the object were taken at equidistant grid points (θ_i, ϕ_i) within the parameter interval [-22,22]x[30,50].

After exposure, the MBR (minimum bounding rectangle) of the figure in the input image was determined and registered onto a 128x128 pixel wide window (resp. a 128x256 window for stereo pairs). Finally, the resulting 128x128 (128x256) image vectors were brightness normalized.

For each figure, images were taken at a equidistant grid (2 degrees in each direction) within the parameter interval [-22,22]x[30,50], thus resulting in a total of 231 mono and 231 stereo image vectors, respectively. The extreme views together with the first eigenvectors of the Garfield mono and stereo image set can be seen in Fig. 4.

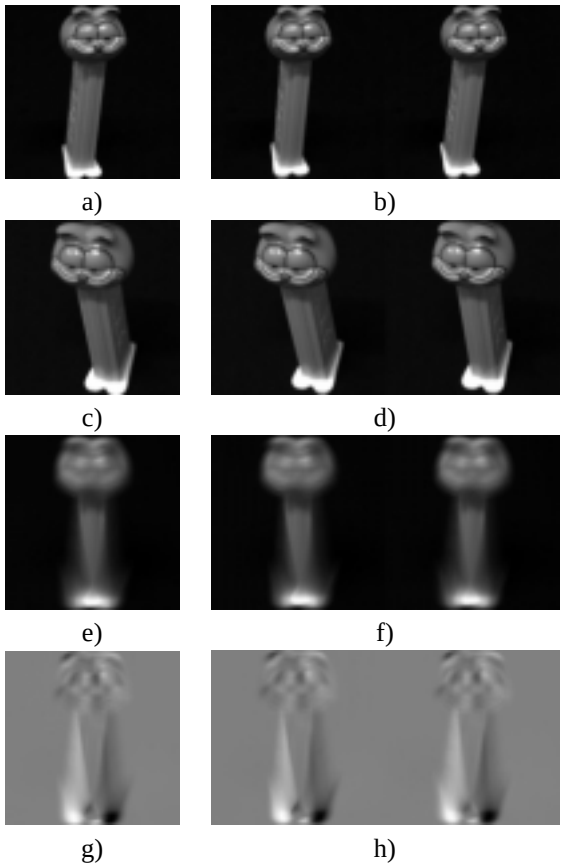
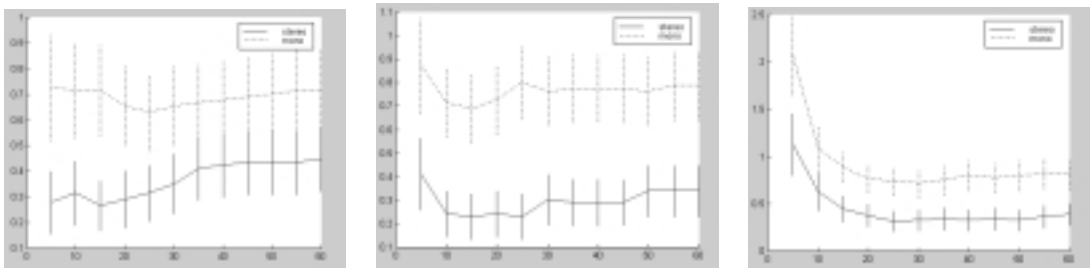


Figure 4: Training set and eigenspace model for Garfield for the mono (left) and stereo (right) approach. a) - b): extreme view at pose (-22,30). c) - d): extreme view at pose (22,50). e) - f): mean image vector. g) - h): first eigenvector.

The training set used in constructing the eigenspace model was obtained by subsampling the original set in increments of 4 degrees, i.e. by discarding every second parameter line; the remaining images were assigned to the test set. A discrete representation of the manifold (a so called reference set) was then built by projecting the training set into the eigenspace, constructing the 2-dimensional parametric manifold by means of bivariate cubic spline-interpolation and resampling it every 2 degrees along each parameter line. The resulting reference set *R* thus consisted of 231 vectors $\{\mathbf{r}_1, .., \mathbf{r}_{231}\}$ (an example of such a parametric manifold can be seen in Fig. 1). For each of the three figures, both a mono and a stereo eigenspace model and associated reference set were built as outlined above.



a) b) c)

Figure 5: Avg. pose estimation errors in degrees for the mono (dashed) and stereo (solid) approach for the Garfield a), Kemal b) and Fozzy c) data set. The pose errors are plotted vs. the dimension of the eigenspace. The vertical lines indicate the standard deviation (scaled by a factor 0.5).

The pose estimation error δ_i for an input image vector $\mathbf{m}_i$ with pose parameters (θ_i, ϕ_i) and eigenspace representation $\mathbf{v}_i$ was calculated as the distance to the parameter values associated with its nearest neighbor in the reference set *R*, $\mathbf{r}_j$:

$$\delta_i = |\theta_j - \theta_i| + |\phi_j - \phi_i|, \tag{4}$$

whereby $j = argmin_k \|\mathbf{r}_k - \mathbf{v}_i\|, k \in \{1..231\}$.

As can be seen from Fig. 5, the stereo eigenspace performed consistently better than its mono counterpart on all three data sets (numerical results for the Garfield set are also given in Table 1). The experiments were performed for several different eigenspace dimensions; the best results for both approaches were typically obtained for an eigenspace dimension of 20 or 25.

As a second experiment, we built two separate mono-eigenspaces (including the discrete manifolds) for the left and right images of the Garfield stereo pairs; these will be referred to as left (LSE) and right (RSE) stereo eigenspace, respectively. A pose estimate for a stereo pair was then obtained by dividing the pair into its left and right subimage and averaging the pose estimates obtained by the LSE and the RSE. As can be seen from the results given in Table 1, this approach performs better than the simple mono-approach, but still not as good as the stereo eigenspace. Note that a beneficial side effect of averaging is the reduction in the variance of the positional error. A distinct disadvantage of the averaging approach is, however, that it has effectively twice the memory and run-time requirements of the mono approach.

4 Conclusion

We have introduced the concept of parametric stereo eigenspaces, in which each object pose relative to the observer is described by two images taken from slightly different positions with fixed relative position and orientation. These images are concatenated to form a single stereo image vector before they are projected into eigenspace. Experimental results indicate that the stereo approach leads to increased accuracy of the estimated pose parameters.

The results given in this paper are preliminary and will have to be elaborated. In particular, we intend to conduct experiments on a larger object data base (including different kinds of objects and additional object poses) and with other representations of the parametric manifold (e.g., Radial Basis Function Neural Networks).

dim	mono		stereo		average	
	avg	stdev	avg	stdev	avg	stdev
5	0.72	1.65	0.27	0.98	0.52	1.02
10	0.71	1.48	0.31	1.01	0.43	0.75
15	0.71	1.41	0.27	0.78	0.44	0.87
20	0.65	1.23	0.29	0.89	0.50	0.95
25	0.63	1.21	0.32	0.88	0.51	0.89
30	0.65	1.23	0.35	0.93	0.53	0.86
35	0.67	1.21	0.41	0.99	0.54	0.85
40	0.68	1.22	0.42	1.01	0.53	0.90

Table 1: Positional error in degrees and its standard deviation for the Garfield data set. These quantities are given for the mono, stereo and averaging approach for eigenspace dimensions in range [5..40]. For all three approaches, the optimal number of eigenvectors lies between 15 and 25.

References

[1] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, second edition, 1990.

[2] H. Hotteling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933.

[3] S. Mukherjee and S. K. Nayar. Optimal RBF networks for visual learning. In *Proc. of IEEE International Conference on Computer Vision*, pages 794–800, 1995.

[4] H. Murase and S. K. Nayar. Illumination planning for object recognition using parametric eigenspaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(12):1219–1227, December 1994.

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

[6] S. K. Nayar, S. A. Nene, and H. Murase. Subspace methods for robot vision. *IEEE Trans. Robotics and Automation*, 12(5):750–758, October 1996.

[7] M. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 4(1):71–86, 1991.