Manifold Pursuit: A New Approach to Appearance Based Recognition

Amnon Shashua *
Stanford University
CS Department
Stanford, CA 94305, USA
shashua@cs.huji.ac.il

Anat Levin *
Stanford University
CS Department
Stanford, CA 94305, USA
alevin@cs.huji.ac.il

Shai Avidan
Interdisciplinary Center
CS Department
Herzliya, Israel
avidan@idc.ac.il

Abstract

*Manifold Pursuit (***MP***) extends Principal Component Analysis to be invariant to a desired group of image-plane transformations of an ensemble of un-aligned images.*

We derive a simple technique for projecting a mis-aligned target image onto the linear subspace defined by the superpositions of a collection of model images. We show that it is possible to generate a fixed projection matrix which would separate the projected image into the aligned projected target and a residual image which accounts for the mis-alignment. An iterative procedure is then introduced for eliminating the residual image and leaving the correct aligned projected target image.

Taken together, we demonstrate a simple and effective technique for obtaining invariance to image-plane transformations within a linear dimensionality reduction approach.

1 Introduction

The "appearance based" paradigm in visual recognition aims at capturing the statistical regularities and redundancies shared by a set of images. It uses the notion of dimensionality reduction of an input image space, in a Linear Coding style, to achieve this goal. Experience shows that for certain applications, and if certain conditions are met, the Linear Coding dimensionality reduction can achieve impressive performance with a very simple and computationally efficient machinery (cf. [19, 15, 5]).

In its most general form, one would like to represent a target image $t(\boldsymbol{x})$ as a linear superposition of basis images $\psi_i(\boldsymbol{x})$:

$$t(\boldsymbol{x}) = \sum_i \lambda_i \psi_i(\boldsymbol{x})$$

where $\boldsymbol{x}$ varies over the two-dimensional plane. The spectrum of proposed methods in this context differ in the way the basis images $\psi_i()$ are recovered from a given set of model images $\phi_1(), ..., \phi_k()$. These include Principle Component Analysis [16, 18, 19], Independent Component Analysis [7, 3], Projection Pursuit [13], Radial Basis functions [10], Factorial coding [2, 11] — for a recent review see [8].

Principal Component Analysis (**PCA**) is commonly used in this context. It assumes that the set of model (and target) images lie in a low dimensional subspace spanned by the eigenvectors (the basis images) of the correlation matrix of the model images.

The strength of **PCA** comes from its efficient computational mechanism, the fact that it is well understood, and from its general applicability. For example, in Vision applications it has been used for the representation and recognition of faces [18, 19], recognition of 3D objects under varying pose [15], tracking of deformable objects [5] and for representations of 3D range data of heads [1].

However, dimension-reducing techniques, like **PCA**, are sensitive to image plane transformations. Consider for example Fig. 1. Five images of correctly aligned and scaled human faces were taken (data set courtesy of the Yale U. face data-base). We then introduce small translations (order of few pixels) to each of the images, and then attempt to reconstruct one of the images (in its original location) as a superposition of the five translated model images. One can clearly see that the lack of invariance is a serious impediment to the representation of a class of objects.

Our goal is to make **PCA** invariant to image-plane transformations, while maintaining the clarity and spirit of **PCA** and without resorting to a complex algorithm.

Previous approaches to the problem include *EigenTracking* [5] that modifies optic-flow equations to work with **PCA**, Tangent-Distance [17] and its multi-resolution extension [20] where the distance between two images is replaced by the distance between the tangent to the image manifolds, and probabilistic approach [9, 14] that separate images into

*The permanent address of A.S and A.L is The School of CS and Engineering, The Hebrew University, Jerusalem , Israel

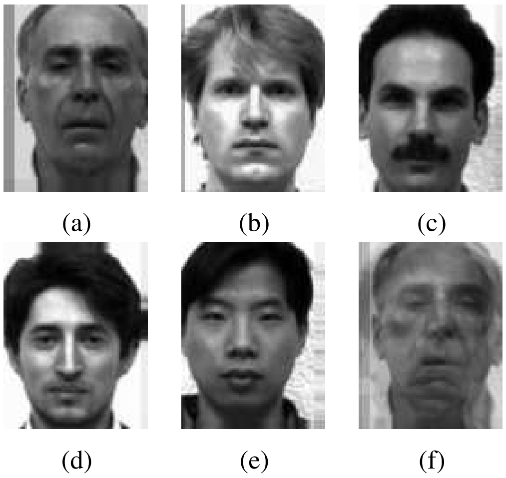

Figure 1. Image Coding using **PCA** depends on correct alignment. (a)-(e) The five images are misaligned by no more than 5 pixels. (f) Projecting the original first image (without translation) on the eigenspace result in the ghost effect seen here.

appearance and deformation.

Our approach to the problem is two fold. First, is to increase the dimensionality of the representation (number of basis images $\psi()$), but in a tightly controlled manner. We show that for small transformations the variability of the input model images live in a linear subspace whose dimension is bounded: for the group of translations the bound is 3 times the dimension of the ensemble without the invariance property, and for general affine transformations the bound is 7. Second, since for large transformations the variability space is no longer embedded in a linear sub-space, we derive a method we call "Manifold Pursuit" (**MP**) for projecting a target image onto the non-linear manifold.

2 Manifold Pursuit

We shortly describe **PCA** as an introduction to **MP**.

2.1 A short introduction to PCA

Let $A = [\phi_1, ..., \phi_k]$ be an $n \times k$ matrix whose columns are the model images (each spread as a vector, where n is the number of pixels of the image). Let $A = UDV^\top$ be the Singular Value Decomposition (SVD) of A, i.e., the columns of U are the eigenvectors of $AA^\top$ and the columns of V are the eigenvectors of $A^\top A$, and D is a diagonal matrix containing the singular values. Near zero singular values correspond to eigenvectors of $AA^\top$ that represent near zero variability of the data (columns of A), hence can be discarded. As a result, the column space of U spans a linear sub-space that best represents the statistical variability of the data. The projection matrix $UU^\top$ projects any target image $t()$ (spread as a vector $\boldsymbol{t}$) onto the linear subspace: $\hat{\boldsymbol{t}} = UU^\top \boldsymbol{t}$. The columns of U are called the Principle Components of the ensemble, and in the context of Vision applications are called "eigenimages". The quality of the reprojection depends a great deal on the assumption that both images of the database and the target image are aligned. Violating this assumption degrades the quality of reprojection. We introduce **MP** to relax this constraint.

2.2 Manifold Pursuit

Let $\phi_1(\boldsymbol{x}), ..., \phi_k(\boldsymbol{x})$ be a set of model images where the index vector $\boldsymbol{x}$ is two dimensional $\boldsymbol{x} = (x_1, x_2)$. An image $t(\boldsymbol{x})$ belongs to the manifold expressed by superpositions of the model images under some group of transformations if the following holds:

$$t(\boldsymbol{x}) = \sum_{i=1}^{k} \lambda_i \phi_i(\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha}^i)) \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$ is a vector function $\boldsymbol{f} = (f_1, f_2)$ representing the group of transformations acting on $\boldsymbol{x}$ with a set of parameters denoted by the vector $\boldsymbol{\alpha}$. For example, in case the invariance we desire is image-plane translation, then $\boldsymbol{f}() = (\alpha^1, \alpha^2)$; and when we desire invariance under affine transformations, then $\boldsymbol{f}() = (a + bx_1 + cx_2, d + ex_1 + fx_2)$ where $\boldsymbol{\alpha} = (a, b, c, d, e, f)$.

The transformation $\boldsymbol{f}()$ is applied independently to every model image, i.e., each model image $\phi_i()$ can undergo an arbitrary transformation (represented by the choice of $\boldsymbol{\alpha}^i$) throughout the linear superposition. Thus, the parameters that are relevant for the representation of the target image $t(\boldsymbol{x})$ are the scalars λ_i and vectors $\boldsymbol{\alpha}^i$.

Assuming small transformations, i.e., $\boldsymbol{\alpha}_i$ are infinitesimal, and considering the first-order Taylor expansion of the right-hand side of eqn. 1 we have that

$$\phi(\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})) \approx \phi(\boldsymbol{x}) + \boldsymbol{\alpha}^\top \phi_{\boldsymbol{\alpha}},$$

where $\phi()$ is a model image and $\phi_\alpha \equiv \frac{\partial \phi}{\partial \boldsymbol{\alpha}}$ is the vector of partial derivatives $(\phi_{\alpha^1}, ..., \phi_{\alpha^p})$, and where

$$\phi_{\alpha^j} = \phi_{f_1} f_{1_{\alpha^j}} + \phi_{f_2} f_{2_{\alpha^j}}$$

For example, for the group of translations the expansion becomes:

$$\phi(\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})) = \phi(x) + \alpha^1 \phi_{x_1} + \alpha^2 \phi_{x_2}, \tag{2}$$

and for the affine group the expansion becomes:

$$\begin{aligned} \phi(\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})) \quad = \quad & \phi(x) + a\phi_{x_1} + bx_1\phi_{x_1} + \\ & cx_2\phi_{x_1} + d\phi_{x_2} + \\ & ex_1\phi_{x_2} + fx_2\phi_{x_2}. \end{aligned} \tag{3}$$

Therefore, for the invariance under pure translation, instead of applying **PCA** on the original matrix $n \times k$ whose columns are the model images ϕ_i (spread as vectors), we should apply **PCA** on an $n \times 3k$ matrix A:

$$A = [\phi_1, \phi_{1_{x_1}}, \phi_{1_{x_2}}, ..., \phi_k, \phi_{k_{x_1}}, \phi_{k_{x_2}}], \tag{4}$$

and for the affine group the $n \times 7k$ matrix:

$$A = [..., \phi_i, \phi_{x_1}, x_1\phi_{x_1}, x_2\phi_{x_1}, \phi_{x_2}, x_1\phi_{x_2}, x_2\phi_{x_2}, ...], \tag{5}$$

The eigenimages of A (eigenvectors of $AA^\top$) span the subspace of model images under the invariance of small image-plane transformations. In effect, the small transformation assumption allows us to represent the manifold of variability of *each* model image with the tangent to the manifold represented by the superposition of functions of first-order partial derivatives of the model image. The only price we pay is to enlarge the set of basis images (the principle components) required for representing the object class under the desired invariance group.

We handle large transformations using Newton iterations and "image warping"within a coarse-to-fine framework. In each Newton iteration we project the target image onto the linear subspace represented by the principle components of A, then we modify (warp) the model images and recompute their derivatives. This procedure is guaranteed to converge to a local minima. To help converging to a global minima the same procedure is performed in a coarse-to-fine manner, using a Gaussian Pyramid [6] of the basis images.

3 The Case of Aligned Model Images

In the case of un-aligned model images we need to re-evaluate the projection matrix in each iteration — making the proposed technique computationally expensive. Now we consider a different variant of the same problem: we assume the model images are aligned but allow the target image to be mis-aligned. This might be useful in the case of face detection followed by identification, where detection module might find the head position up to some translational/rotational ambiguity and thus the target image will not be aligned with the model images.

The advantage of assuming an aligned model set is that, as shown next, the projection matrix will need to be evaluated only once thereby reducing the computational cost considerably. We consider the following problem:

$$t(\boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha}_i)) = \sum_{i=1}^{k} \lambda_i \phi_i(\boldsymbol{x}) \tag{6}$$

where $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$ is defined as in the previous section. For simplicity we will start with a translational model $\boldsymbol{f}() = (\alpha^1, \alpha^2)$. From the arguments of the previous section, the translated target image is spanned by the model images and their derivatives (as a first approximation):

$$\hat{\boldsymbol{t}} = A(A^\top A)^{-1} A^\top \boldsymbol{t}$$

where $\boldsymbol{t}$ is the target image $t()$ spread as a vector, and A is the $n \times 3k$ matrix defined below:

$$A = [\phi_1, ..., \phi_k, \phi_{1_{x_1}}, ..., \phi_{k_{x_1}}, \phi_{1_{x_2}}, ...,, \phi_{k_{x_2}}]. \tag{7}$$

The vector $\hat{\boldsymbol{t}}$ is the projection of the un-aligned target image onto the subspace spanned by the model images and their first partial derivatives. Let $\boldsymbol{y} = (A^\top A)^{-1} A^\top \boldsymbol{t}$ be the coefficients of representing $\hat{\boldsymbol{t}}$ in the new basis spanned by the columns of A, and let $\hat{\boldsymbol{t}} = \boldsymbol{t}' + \boldsymbol{t}''$ where

$$\boldsymbol{t}' = A_{1..k} y_{1..k} \in span\{\phi_1, ..., \phi_k\}$$

and

$$\boldsymbol{t}'' = A_{k+1..3k} y_{k+1..3k} \in span\{\phi_{1_{x_1}}, ..., \phi_{k_{x_1}}, \phi_{1_{x_2}}, ...,, \phi_{k_{x_2}}\},$$

where $A_{i..j}$ denote the sub-matrix consisting of columns i through j, and $y_{i..j}$ denote the sub-vector consisting of entries i through j. Since A has a linearly independent column space, $\boldsymbol{t}', \boldsymbol{t}''$ are uniquely defined. We have therefore a residual image $\boldsymbol{t}''$ and an image $\boldsymbol{t}'$ in the subspace spanned by the original (aligned) model images. Let α^1, α^2 be the least-squares optimized translation between the image $t'()$ and the original target image $t()$ — for example using the gradient-based approach described in [4]. We will use α^1, α^2 to warp $t()$ towards to $t'()$ and project the warped image again onto the subspace spanned by the columns of A — thus gradually reducing the residual image $t''()$.

The projection matrix remains fixed, and so the process involves repetitive projections using a fixed projection matrix while finding the "best" image-plane transformation between "one half" of the projection and the original target image.

There is no need to utilize the entire projection matrix as we are interested only in $\boldsymbol{t}'$. Thus, let A' consist of the first k columns of A (i.e., the original aligned model images) and $A'' = [(A^\top A)^{-1} A^\top]_{1..k}$. We then have that $\boldsymbol{t}' = A' A'' \boldsymbol{t}$. Therefore the computational cost for obtaining $\boldsymbol{t}'$ is proportional to the number of model images k rather than the enlarged space ($3k$ for translational model and $7k$ for Affine model).

To summarize, the method for projecting an unaligned target image $t()$ onto the subspace of model images $\phi_1(), ..., \phi_k()$ is as follows:

1. Let $A' = [\phi_1, ..., \phi_k]$ be the matrix whose columns are the model images spread as vectors, and let A be defined as in eqn. 7 (i.e., the matrix whose columns are the model images and their first partial derivatives). Let $A'' = [(A^\top A)^{-1} A^\top]_{1..k}$, i.e., the first k columns of $(A^\top A)^{-1} A^\top$.

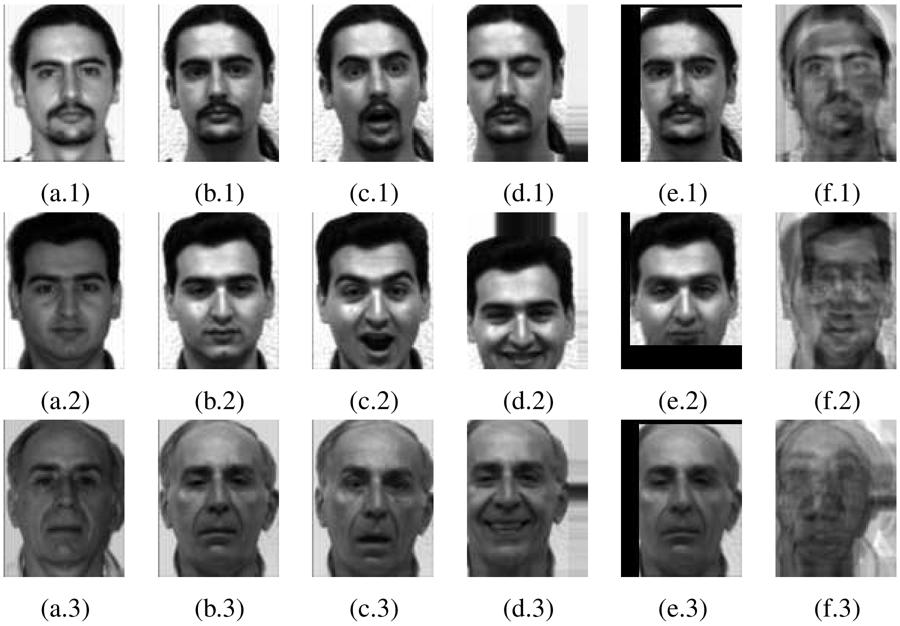

Figure 2. A sample of projection results. The first three columns show the three (aligned) model images of three persons (out of 15 people of the dataset). The variation covers facial expressions and illumination. The fourth column shows a shifted target image and the projected image using **MP** is shown in the fifth column. The sixth column shows, for comparison, the result of projection using **PCA**.

2. Let $\boldsymbol{t}' = A'A''\boldsymbol{t}$.
3. Find α^1, α^2 which minimize $\sum_{i,j}(t(i,j) - t'(i - \alpha^1, j - \alpha^2)^2$.
4. Warp the image $t()$ with α^1, α^2 and go to Step 2. The process ends when the residual displacement α^1, α^2 is sufficiently small.

The process above is implemented within a coarse-to-fine framework (as explained in the previous section) where in each level of the pyramid the procedure above is applied.

4 Implementation Results

We applied **MP** for identifying frontal images of human faces under variability of facial expressions and illumination conditions. We used the Yale U. image-set consisting of aligned frontal human faces covering 9 images per person over 15 distinct people. Fig. 2 shows a sample of the images in the data set.

We selected three images per person as the model images, i.e., an *aligned* target image would be matched to a model (a person) if the distance to its projection onto the linear subspace spanned by the three model images is the smallest over all the 15 models. The first three columns of Fig. 2 show the model images of three persons in the data set. The remaining 6 images per person formed the testing set for our experiment.

Each image of the test set was then shifted (translated) by a measure of up to 20% of the image size (approximately 15 pixels in the horizontal and vertical axes) and then projected onto the three model images (per person). The fourth column of Fig. 2 shows a test image for each of the three persons in the figure. The fifth column shows the reconstructed image (the projection) using **MP**. Note that the shift was recovered during the projection process. For comparison, the sixth column shows the projection, using **PCA**, of the test image (of column 4) onto the the model images — note the "ghost" effect due to the mis-alignment of the test image and the model set.

Table 1 compares the identification success over the test images shifted by various magnitudes. The rows of the table correspond to the range of the shift applied to the original dataset. The second column shows the percentage of correct identification when using **PCA** and, as expected from the results of the previous figure, the performance degrades rapidly with increasing shift magnitude. The third and fourth columns show the identification performance using **MP** when the model images are unaligned (shifted deliberately in the experiment) and when the model images are aligned (yet the test image is un-aligned). Note for example the fourth column: the identification performance is

Range of Shift	**PCA**	**MP** un-aligned	**MP** aligned
$[-5..5]$	93%	99%	100%
$[-7.5..7.5]$	77%	98%	100%
$[-10..10]$	67%	96%	100%
$[-12.5..12.5]$	54%	91%	98%
$[-15..15]$	50%	85%	96%

Table 1. Identification statistics over the test images shifted by various magnitudes. The rows of the table correspond to the range of the shift (in pixels) applied to the original dataset. The columns correspond to the use of **PCA**, **MP** using an un-aligned model set of images, and **MP** with an aligned model set. Note that the identification performance degrades rapidly with increasing shift magnitude (second column) compared to the largely invariant performance to shift magnitude with **MP** (fourth column).

largely invariant to the magnitude of shift.

5 Summary

Manifold Pursuit is a simple and efficient method for representing an object class using linear dimensionality reduction methodology while maintaining invariance over a desired group of transformations. The method consists of two components: (i) enlarging the set of principle components by including functions of first-order derivatives of the model images, and (ii) performing the projection iteratively within a Gaussian Pyramid, thus implementing Newton iterations for projecting the target image onto the non-linear manifold.

We presented two definitions of the problem, the first when the model images are unaligned with each other and the second when the model images are aligned but the target image is un-aligned. The advantages of the second approach is in that the projection engine remains fixed throughout the iterative process thereby introducing a simple and computationally efficient method.

The simplicity of the approach enables the saving of exhaustive search when matching a candidate image region to a model under a group of image-plane transformations.

References

[1] J.J. Atick, P.A. Griffin and N.A. Redlich. Statistical approach to shape-from-shading: deriving 3D face surfaces from single 2D images. In *Neural Computation*. To appear.

[2] H.B. Barlow, T.P. Kaushal and G.J. Mitchison. Finding minimum entropy codes. In *Neural Computation 1(3)*, pages 412-423, 1989.

[3] A.J. Bell and T.J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. In *Neural Computation 7(6)*, pages 1129-1159, 1995.

[4] J.R. Bergen, P. Anandan, K.J. Hanna and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision*, Santa Margherita Ligure, Italy, June 1992.

[5] Michael J. Black and Allan D. Jepson. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. In *Proceedings of the European Conference on Computer Vision*, pages 329–342, Cambridge, England, 1996.

[6] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. In *IEEE Transactions on Communication*, 31(532:540), 1983.

[7] P. Comon. Independent component analysis, a new concept? In *Signal processing 36(3)*, pages 11-20, 1994.

[8] G. Deco and D. Obradovic. An information-theoretic approach to neural computing. In New York:Springer-Verlag.

[9] B. Frey and N. Jojic. Transformed component analysis: joint estimation of spatial transformations and image components. In *International Conference on Computer Vision*, Corfu, 1999.

[10] Federico Girossi and Tommaso Poggio. Networks and the best Approximation Property. In *Biol. Cybern.* 63, pages 169-176, 1990.

[11] Geoffrey E. Hinton, Peter Dayan and Michael Revow. Modeling the manifolds of Images of Handwritten Digits. Submitted to *III on Neural Networks*, January 1996.

[12] B.K.P. Horn and B.G. Schunk. Determining Optical Flow. In *Journal of Artificial Intelligence* 17(185-203), 1981.

[13] N. Intrator. Feature extraction using an unsupervised neural network. In *Neural Computation 4*, pages 98-107, 1992.

[14] N. Jojic, P. Simard and B. Frey. Separating Appearance from Deformation. In *International Conference on Computer Vision*, Vancouber, Canada, 2001.

[15] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. In *International Journal of Computer Vision*, 14(5-24), 1995.

[16] E. Oja. Principal components and linear neural networks. In *Neural Networks 5*, pages 927-935, 1989.

[17] P. Simard, B. Victorri, Y. Le Cun and J. Denker. Tangent Prop — a formalism for specifying selected invariances in an adaptive network. In *Proceedings of the fourth annual conference NIPS*, pages 895-899, Denver, CO., 1991.

[18] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. In *Journal of the Optical Society of America 4*, pages 519-524, 1987.

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

[20] N. Vasconcelos and A. Lippman. Multiresolution tangent distance for affine-invariant classification. In *NIPS, 10*, 1998.