Transformation-Invariant Embedding for Image Analysis

Ali Ghodsi[1], Jiayuan Huang[1], and Dale Schuurmans[2]

[1] School of Computer Science
University of Waterloo
{aghodsib,j9huang}@cs.uwaterloo.ca
[2] Department of Computing Science
University of Alberta
dale@cs.ualberta.ca

Abstract. Dimensionality reduction is an essential aspect of visual processing. Traditionally, linear dimensionality reduction techniques such as principle components analysis have been used to find low dimensional linear subspaces in visual data. However, sub-manifolds in natural data are rarely linear, and consequently many recent techniques have been developed for discovering non-linear manifolds. Prominent among these are *Local Linear Embedding* and *Isomap*. Unfortunately, such techniques currently use a naive appearance model that judges image similarity based solely on Euclidean distance. In visual data, Euclidean distances rarely correspond to a meaningful perceptual difference between nearby images. In this paper, we attempt to improve the quality of manifold inference techniques for visual data by modeling local neighborhoods in terms of natural transformations between images—for example, by allowing image operations that extend simple differences and linear combinations. We introduce the idea of modeling local tangent spaces of the manifold in terms of these richer transformations. Given a local tangent space representation, we then embed data in a lower dimensional coordinate system while preserving reconstruction weights. This leads to improved manifold discovery in natural image sets.

1 Introduction

Recently there has been renewed interest in manifold recovery techniques motivated by the development of efficient algorithms for finding non-linear manifolds in high dimensional data. Isomap [1] and Local Linear Embedding (LLE) [2] are two approaches that have been particularly influential. Historically, two main ideas for discovering low dimensional manifolds in high dimensional data have been to find a mapping from the original space to a lower dimensional space that: (1) preserves pairwise distances (i.e. multidimensional scaling [3]); or (2) preserves mutual linear reconstruction ability (i.e. principle components analysis [4]). In each case, globally optimal solutions are linear manifolds. Interestingly, the more recent methods for manifold discovery, Isomap and LLE, are based on

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 519–530, 2004.

exactly these same two principles, with the generalization that the new methods only seek manifold descriptions that *locally* preserve distances and linear reconstructions. In this way, they avoid recovering linear global solutions [1,2].

There have been many new variants of these ideas [5,6,7,8]. Although these techniques all produce non-linear manifolds in different ways, they are generally based on the core assumption that, in natural data, (1) Euclidean distances locally preserve geodesic distances on the manifold [1], or (2) data objects can be linearly reconstructed from other data points nearby in Euclidean distance [2]. However, these core notions are not universally applicable nor always effective. Particularly in image data it is easy to appreciate the shortcoming of these ideas: For images, weighted linear combinations amount to an awkward transformation whereby source images have their brightness levels adjusted and then are summed directly on top of one another. This is often an unnatural way to capture the image transformations that manifolds are intended to characterize. Figure 1 shows that centered, cropped and normalized target images can be reasonably well reconstructed from likewise aligned source images, but that even a minor shift, rotation or rescaling will quickly limit the ability of this approach to reconstruct a target image. Similarly, measuring Euclidean distances between images can sometimes be a dubious practice, since these distances do not always correspond to meaningful perceptual differences.

Fig. 1. Least squares reconstructions of a target image (far right) from three nearby images (far left). The intermediate (fourth) image shows the best linear reconstruction of the rightmost image from the three leftmost images. First row: original reconstruction. Second row: reconstruction of same image after translations have been applied.

We propose to model manifolds locally by characterizing the local transformations that preserve the invariants they encode. That is, we attempt to characterize those transformations that cause points on the manifold to stay on the manifold. Our approach will be to first characterize the local tangent space around a data object by considering transformations of that object that cause it to stay on (or near) the manifold.

Other work on incorporating natural image transformations to better model visual data has been proposed by [9,10,11,12]. However, this previous work primarily concerns learning mixture models over images rather than sub-manifolds, and most significantly, requires that the image transformations be manually spec-

ified ahead of time, rather than inferred from the data itself. In this paper, we infer local transformations directly from the image data.

Eigentracking [13], also considers affine transformations of a set of preconstructed basis images for an object based on preliminary views. Here we consider a potentially richer set of transformations and simultaneously learn the basis in addition to the transformations and embedding.

2 Local Image Transformations

For images, it is easy to propose simple local transformations that capture natural invariants in image data better than simply averaging nearby images together. Consider a very simple class of transformations based on receptive fields of pixel neighborhoods: Given an $n_1 \times n_2$ image x, imagine transforming it into a nearby image $\tilde{x} = T(x, \theta)$, where for each pixel $\tilde{x}_i \in \tilde{x}$ we determine its value from corresponding nearby pixels in x. Specifically, we determine $\tilde{x}_i$ according to

$$\tilde{x}_i = \theta^\top x_{N(i)} \tag{1}$$

where $N(i)$ denotes the set of neighboring pixels of pixel x_i. Thus $T(\cdot, \theta)$ defines a simple local filter passed over the image, parameterized by a single weight vector θ, as shown in Figure 2.

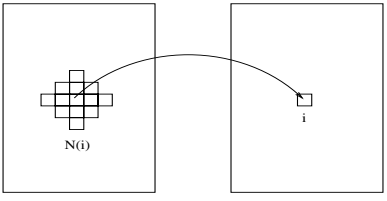

Fig. 2. Illustration of local pixel transformation from the left image to the right

Although this defines a limited class of image transformations, it obviously enhances the image modeling capabilities of weighted image combinations (which are only based on adjusting the brightness level of source images). Many useful types of transformation such as translation, rotation and blurring can be approximated using this simple local transformation. Figure 3 shows that similar images can be much better reconstructed by simple filter transformations rather than merely adjusting brightness levels prior to summing. Here minor translations and appearance changes can be adequately modeled in circumstances where brightness changes fail.

3 Local Tangent Space Modeling

The key to our proposal is to model the local tangent space around high-dimensional data points by a small number of transformations that locally pre-

Fig. 3. Least squares reconstructions of a target image (far right) from three nearby images (left). The intermediate (fourth) image shows the best least squares reconstruction of the rightmost image from the three leftmost images. First row: standard reconstruction. Second row: reconstruction after local transformations.

serve membership in the manifold. Thus, in our approach, a manifold is locally characterized by the invariants it preserves.

We model transformations over the data space by using an operator $T(x, \theta)$ which combines a data object x and a parameter vector θ to produce a transformed object $\tilde{x} = T(x, \theta)$. In general, we will need to assume very little about this operator, but, by making some very simple (and fairly weak) assumptions about the nature of T, we will be able to formulate natural geometric properties that one can preserve in a dimensionality reducing embedding.

First, we assume that T is a *bilinear* operator. That is, T becomes a linear operator on each argument when the other argument is held fixed. Specifically,

$$\begin{aligned} T(ax_1 + bx_2, \theta) &= aT(x_1, \theta) + bT(x_2, \theta) \\ T(x, a\theta_1 + b\theta_2) &= aT(x, \theta_1) + bT(x, \theta_2) \end{aligned} \tag{2}$$

Second, we require the operator to have a local *origin* ω in the second argument that gives an identity map:

$$T(x, \omega) \;=\; x \text{ for all } x \tag{3}$$

With these properties, we can then naturally equate parameterized transformations with tangent vectors as follows. First note that $T(x, \theta) = x + T(x, \delta)$ for $\delta = \theta - \omega$, since by bilinearity we have

$$T(x, \theta) = T(x, \omega + \delta) = T(x, \omega) + T(x, \delta)$$

and also

$$T(x, \omega) = x$$

Thus, we can interpret every transformation of an object x as a vector sum. That is, if $\tilde{x} = T(x, \theta)$ then the difference $\tilde{x} - x$ is just $T(x, \delta)$.

Now imagine transforming a source object x_i to approximate a nearby target object x_j, where both reside on the manifold. The best approximation of x_j by x_i is given by

$$\tilde{x}_{ij} = T(x_i, \tilde{\theta}_{ij})$$

where

$$\tilde{\theta}_{ij} = \arg\min_{\theta} \|x_j - T(x_i, \theta)\|$$

If the approximation error is small, we can claim that the difference vector $\tilde{x}_{ij} - x_i = T(\tilde{\delta}_{ij})$, for $\tilde{\delta}_{ij} = \tilde{\theta}_{ij} - \omega$, is approximately tangent to the manifold at x_i. One thing we would like to preserve is the transformation distance between nearby points. Consider the norm of the difference vector:

$$\|x_i - \tilde{x}_{ij}\| \;=\; \|T(x, \tilde{\delta}_{ij})\| \;=\; \|\tilde{\delta}_{ij}\|\, \|T(x, \tilde{\eta}_{ij})\|$$

where $\tilde{\eta}_{ij} = \tilde{\delta}_{ij}/\|\tilde{\delta}_{ij}\|$. Here $T(x, \tilde{\eta}_{ij})$ gives the direction of the approximate tangent vector at x_i, and $\|\tilde{\delta}_{ij}\|$ gives the coefficient in direction $\tilde{\eta}_{ij}$. This says that $\tilde{x}_{ij}$ is the projection of x_j onto the tangent plane centered at x_i, since $\tilde{x}_{ij} = x_i + \|\tilde{\delta}_{ij}\| T(x, \tilde{\eta}_{ij})$ is the best approximation of x_j in the local tangent space of x_i.

Intuitively, when we embed x_i and $\tilde{x}_{ij}$ in a lower dimensional space, say by a mapping $x_i \mapsto y_i$ and $\tilde{x}_{ij} \mapsto \tilde{y}_{ij}$, we would like to preserve the coefficient:

$$\|y_i - \tilde{y}_{ij}\| \;\approx\; \|\tilde{\delta}_{ij}\|$$

That is, in the lower-dimensional space, the vector $y_i - \tilde{y}_{ij}$ encodes the embedded direction of the transformation, $T(x_i, \tilde{\eta}_{ij})$, and the length $\|y_i - \tilde{y}_{ij}\|$ encodes the coefficient of the transformation, $\|\tilde{\delta}_{ij}\|$.

4 Transformation-Invariant Embedding Algorithm

Consider a set of t vectors, x_i, of dimension n sampled from an underlying manifold. If the manifold is smooth and locally invariant to natural transformations, we should be able to transform nearby points on the manifold to approximate each other. Therefore, in the low dimensional embedding we would like to preserve the ability to reconstruct points from their transformed neighbors. First, to identify the local neighborhood of each data point x_i, we compute the best point-to-point approximations using the local transformation operator described above (as opposed to just using Euclidean distances as proposed in LLE and Isomap). That is, given a target image x_j and a source image x_i, the best approximation of x_j from source x_i is given by

$$\tilde{x}_{ij} = T(x_i, \tilde{\theta}_{ij})$$

where

$$\tilde{\theta}_{ij} = \arg\min_{\theta} \|x_j - T(x_i, \theta)\|$$

Given these quantities, the neighborhood of an image x_j can then be approximated by selecting the K nearest neighbors x_i according to the K best approximations among the transformed reconstructions $\tilde{x}_{ij}$.

Second, to characterize the structure of the local neighborhood, we re-express each data point x_j in terms of its K nearest reconstructions $\tilde{x}_{ij}$. Consider a

particular image x_j with K nearest neighbors $\tilde{x}_{ij}$ and reconstruction weights w_{ij}. The reconstruction error can be written as:

$$\varepsilon_j(w_j) = \left\| x_j - \sum_{i=1}^{K} w_{ij} \tilde{x_{ij}} \right\|^2$$

where w_j is the vector of reconstruction weights for an image x_j in terms of its neighbors. Note that each data point x_j is reconstructed independently. That is, we can recover each set of weights separately by solving a system of n linear equations in K unknowns. This can be expressed in a standard matrix form

$$\varepsilon_j(w_j) = \left\| \chi_j w_j - N_j w_j \right\|^2 = \left\| (\chi_j - N_j) w_j \right\|^2 = w_j^T G_j w_j$$

where χ_j is the matrix of columns x_j repeated K times, N_j is the matrix of columns of K nearest reconstructions $\tilde{x}_{ij}$ of x_j, and $G_j = (\chi_j - N_j)^T (\chi_j - N_j)$.

Note that, as with LLE, we wish to preserve scale and translation invariance in the local manifold characterization, and therefore we impose the additional constraint that the reconstruction weights w_j of each point x_j from its transformed neighbors $\tilde{x}_{ij}$ sums to one. That is, $\sum_i w_{ij} = 1$ for all j. The rationale for this constraint is that we would like the reconstruction weights to be invariant under the mapping from the neighborhood to the global manifold coordinates, which can be shown to hold if and only if all rows of the weight matrix sum to one [2]. Therefore, imposing the extra constraint ensures that the reconstruction holds equally well in both high dimensional and low dimensional spaces. To show that the resulting constrained least squares problem can still be solved in closed form, introduce a Lagrange multiplier λ and let e be a column vector of ones, obtaining

$$L(w, \lambda) = w^T G w + \lambda (w^T e - e)$$
$$\frac{dL}{dw} = 2Gw + \lambda e = 0$$
$$Gw = Ce$$

In practice, we can solve this with C set arbitrarily to 1 and then rescale so w sums to 1.

Finally, we need to embed the orginal points x_j in the lower dimensional coordinate system by assigning them coordinates y_j. Here we follow the same approach as LLE and choose the d dimensional vectors y_j to minimize the embedding cost function

$$\Phi(Y) = \sum_{j=1}^{t} \left\| y_j - \sum_{i=1}^{t} w_{ij} y_i \right\|^2$$

This ensures that we maintain the reconstruction ability in the coordinate system of the lower dimensional manifold. To solve for these coordinates, re-express the cost function in a standard matrix form

$$\Phi(Y) = \sum_{j=1}^{t} \left\| Y I_j - Y w_j \right\|^2$$

where I_j is the j^{th} column of the identity matrix, and w_j is the j^{th} column of W. Then we obtain

$$\min_Y \sum_{j=1}^{t} \left\| YI_j - Yw_j \right\|^2 = \min_Y \ trace(YMY^T)$$

where $M = (I - W)^T(I - W)$. As observed in [2] the solution for Y can have an arbitrary origin and orientation, and thus to make the problem well-posed, these two degrees of freedom must be removed. Requiring the coordinates to be centered on the origin ($\sum_j y_j = 0$), and constraining the embedding vectors to have unit covariance ($Y^TY = I$), removes the first and second degrees of freedom respectively. So the cost function must be optimized subject to additional constraints. Considering only the second constraint for the time being, we find that

$$L(Y, \lambda) = YMY^T + \lambda(YY^T - (N-1)I)$$

$$\frac{dL}{dY} = 2MY^T + 2\lambda Y = 0$$

$$MY^T = \lambda Y^T$$

Thus L is minimized when the columns of Y^T (rows of Y) are the eigenvectors associated with the lowest eigenvalues of M. Discarding the eigenvector associated with eigenvalue 0 satisfies the first constraint.

5 Experimental Results

We present experimental results on face image data. The first two experiments attempt to illustrate the general advantages of the proposed technique, Transformation Invariant Embedding (TIE), for discovering smooth manifolds, at least in simple image analysis problems. A subsequent experiment attempts to show some of the advantages for TIE in a face recognition setting. In all experiments we use the transformation operator on images (1) that was described in Section 2.

Our first experiment is on translated versions of a single face image, as shown in Figure 4. Although the data set is high dimensional (the images are comprised of many pixels), there is clearly a one dimensional manifold that characterizes the image set. Figure 4 shows the result of running LLE and TIE on the original data set shown at the top. The results show that the 1-dimensional manifold discovered by LLE is inferior to that discovered by TIE, which had no problem tracking the vertical shift in the image set.

We then conducted an experiment on a database of rotating face images. Figure 5 shows the two-dimensional manifold discovered by LLE, whereas Figure 5 shows the two-dimensional manifold recovered by TIE. In both cases, the first dimension (top) captured the rotation angle of the images, although once again LLE's result is not as good as TIE's. Interestingly, TIE (and to a lesser extent LLE) learned to distinguish frontal from profile views in its second dimension.

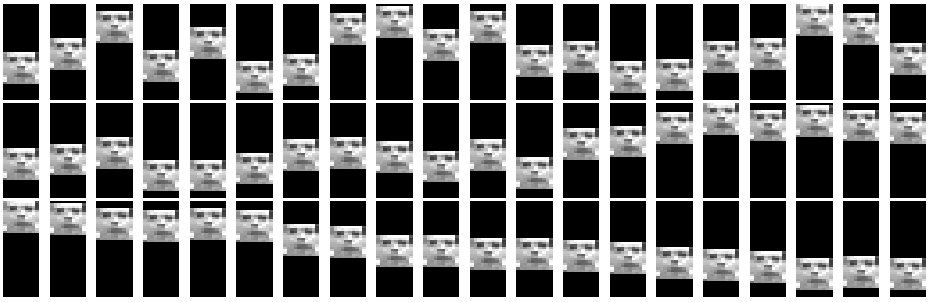

Fig. 4. Top: Original data. Middle: 1-dimensional manifold discovered by LLE. Bottom: 1-dimensional manifold discovered by TIE. (Images are sorted by the 1-dimensional y-coordinate values assigned by LLE and TIE respectively.)

Fig. 5. Two-dimensional manifold discovered by LLE. Top two rows show first dimension, bottom two rows show second dimension.

Fig. 6. Two-dimensional manifold discovered by TIE. Top two rows show first dimension, bottom two rows show second dimension. Note: first dimension captures rotation, whereas second captures frontal views versus side views.

Fig. 7. 105 rotated face images of 7 subjects

Finally, we conducted an experiment on a database of face images that contains 105 face images of 7 subjects which includes variations in both pose, and lighting (see Figure 7). The original data space was embedded into three dimensional subspaces. Figures 8 and 11 show the first dimension discovered by LLE and TIE respectively. Similarly, Figures 9 and 12 show the second dimension for LLE and TIE; and Figures 10 and 5 show the third dimension.

Note that for TIE the first (Figure 11) and second (Figure 12) dimensions corespond to rotation and frontal and profile views, whereas TIE essentially learned to distinguish faces in its third dimension (Figure 5). Here, two individuals were confused by TIE, whereas the other subjects were separated very well.

The corresponding results for LLE are clearly inferior in each case. Figures 8, 9 and 10 illustrates that LLE failed to discover smooth rotations, frontal versus side views, and identity.

6 Conclusion

In many image analysis problems, we know in advance that the data will incorporate different types of transformations. We introduce a way to make standard manifold learning methods such as LLE invariant to transformations in the input. This is achieved by modeling the local tangent space around high-dimensional data points by a small number of transformations that locally preserve membership in the manifold. Thus, in our approach, a manifold is locally characterized by the invariants it preserves.

Fig. 8. First dimension of the three-dimensional manifold discovered by LLE

Fig. 9. Second dimension of the three-dimensional manifold discovered by LLE

Fig. 10. Third dimension of the three-dimensional manifold discovered by LLE

Fig. 11. First dimension of the three-dimensional manifold discovered by TIE

Fig. 12. Second dimension of the three-dimensional manifold discovered by TIE

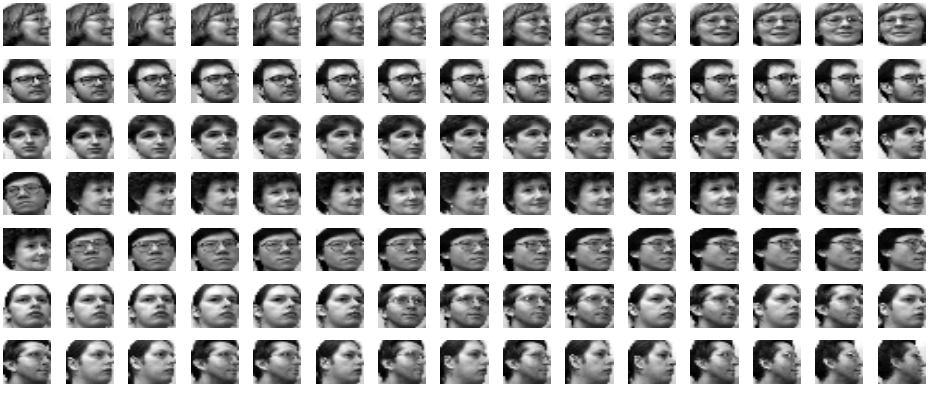

Fig. 13. Third dimension of the three-dimensional manifold discovered by TIE

We model transformations over the data space by using a bilinear operator which produce a transformed object, and show that by making this fairly weak assumption about the nature of operator, we will be able to formulate natural geometric properties that one can preserve in a dimensionality reducing embedding.

Although our basic approach is general, we focused on the special case of modeling manifolds in natural image data with emphasis on face recognition data. Here the proposed a simple local transformations capture natural invariants in the image data better than simply averaging nearby images together. Although we have focused solely on facial rotation and translation as the basic invariants we have been attempting to capture, clearly other types of transformations, such as warping, and out of plane rotation, are further phenonenon one may with to capture with these techniques.

References

1. J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
2. L. Saul and S. Roweis. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. *JMLR*, 2003.
3. T. Cox and M. Cox. *Multidimensional Scaling*. Chapman Hall, 2nd edition, 2001.
4. I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.
5. G. Lebanon. Learning Riemannian metrics. In *Proceedings UAI*, 2003.
6. M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings NIPS*, 2001.
7. Y.-W. Teh and S. Roweis. Automatic alignment of local representations. In *Proceedings NIPS*, 2002.
8. G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Proc. NIPS*, 2002.
9. B. Frey and N. Jojic. Estimating mixture models of images and inferring spatial transformations using the em algorithm. In *Proceedings CVPR*, 1999.
10. B. Frey and N. Jojic. Transformed component analysis: joint estimation of spatial transformations and image components. In *Proceedings ICCV*, volume 2, pages 1190–1196, 1999.
11. B. Frey and N. Jojic. Transformation-invariant clustering and dimensionality reduction using em. *To appear in IEEE PAMI*, 2003.
12. P. Simard, Y. Le Cum, and J. Denker. Efficient pattern recognition using a new transformation distance. In *Proceedings NIPS 5*, 1993.
13. M. Black and A. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1), 1998.