Semi-Supervised Nonlinear Dimensionality Reduction

Xin Yang XINYANG@CSE.PSU.EDU
Haoying Fu HFU@CSE.PSU.EDU
Hongyuan Zha ZHA@CSE.PSU.EDU
Jesse Barlow BARLOW@CSE.PSU.EDU
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA

Abstract

The problem of nonlinear dimensionality reduction is considered. We focus on problems where prior information is available, namely, semi-supervised dimensionality reduction. It is shown that basic nonlinear dimensionality reduction algorithms, such as Locally Linear Embedding (LLE), Isometric feature mapping (ISOMAP), and Local Tangent Space Alignment (LTSA), can be modified by taking into account prior information on exact mapping of certain data points. The sensitivity analysis of our algorithms shows that prior information will improve stability of the solution. We also give some insight on what kind of prior information best improves the solution. We demonstrate the usefulness of our algorithm by synthetic and real life examples.

1. INTRODUCTION

With the development of science, more and more areas of science need to deal with large volumes of high-dimensional data, such as human gene distributions, global climate patterns, etc. In many application fields, high dimensional data need to be analyzed and/or visualized. This leads to the research of dimensional reduction: to find a meaningful low-dimensional manifold from the high-dimensional data. Traditionally, multidimensional scaling (MDS) (Hastie et al., 2001) and principal component analysis (PCA) (Hastie et al., 2001) have been used for dimensionality reduction.MDS and PCA perform well if the input data lie on or are close to a linear subspace, but are not designed to discover nonlinear structures, and often fail to do so.

In many real world applications, data samples lying in a high dimensional ambient space can be modeled by very low dimensional nonlinear manifolds. For example, in the problem of moving object detection and tracking, the dimensionality of frames from a video sequence are usually considered to be the number of pixels of the frames, which can be very high. However, if the video sequence shows a moving object, then the coordinates of the moving object in each frame bear much of the information in that frame, therefore, the frames actually lie on a low dimensional nonlinear manifold. Recently, there have been much research effort on nonlinear dimensionality reduction. For example, the Locally Linear Embedding (LLE) (Roweis & Saul, 2000),(Saul & Roweis, 2003) algorithm computes a global coordinate system of low dimension by finding a low-dimensional space that best preserves the neighborhood of the input data points. The ISOMAP (Tenebaum et al., 2000) approach seeks to preserve the geodesic manifold distance rather than the Euclidean distance between all pairs of data points. The Local Tangent Space Alignment (LTSA) (Zhang & Zha, 2004),(Zha & Zhang, 2005) method constructs an approximation for the tangent space at each data point, and align these tangent spaces to give the global coordinates of the data points. Weinberger et al (Weinberger et al., 2005) proposed using semi-definite programming and kernel matrix factorization to maximize the variance in feature space while preserving the distance and angles between nearest neighbors.

Classical methods, such as LLE, ISOMAP, and LTSA are all unsupervised learning algorithms, that is, they assume no prior information on the input data. Furthermore, these algorithms do not always yield low dimensional coordinates that bear any physical meaning. Here we extend these algorithms to take into

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006.

account prior information. Prior information can be obtained from experts on the subject of interest and/or by performing experiments. For example, in moving object tracking, the coordinates of the object in certain frames can be determined manually, and can be used as prior information. We consider prior information in the form of on-manifold coordinates of certain data samples. We consider both exact and inexact prior information. We call the new algorithms Semi-Supervised LLE (SS-LLE), Semi-Supervised ISOMAP (SS-ISOMAP), and Semi-Supervised LTSA (SS-LTSA). Assuming the prior information has a physical meaning, then our semi-supervised algorithms yield global low dimensional coordinates that bear the same physical meaning.

The rest of the paper is organized as follows. In §2, we give a brief description of the LLE, ISOMAP, and LTSA algorithms. In §3, we show how to extend the basic LLE, ISOMAP, and LTSA algorithms such that they can handle exact prior information. In §4, we present a sensitivity analysis of our algorithms, which shows that prior information will improve stability of the solution, and gives insight on what kind of prior information best improves the solution. We discuss how to deal with inexact prior information in §5. In §6, we apply our algorithms to a synthetic dataset and a real life dataset that was used for motion tracking, conclusions are made in §7.

2. THE BASIC LLE, LTSA, AND ISOMAP ALGORITHMS

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a set of n real-valued vectors, where each $\mathbf{x}_i \in \mathcal{R}^D$ is sampled from some underlying nonlinear manifold given as

$$\mathbf{x}_i = \mathbf{f}(\mathbf{y}_i) + \mathbf{u}_i, \quad i = 1, 2, \cdots, n. \tag{1}$$

Here $\mathbf{y}_i \in \mathcal{R}^d$ represents the sought after low dimensional feature vector of $\mathbf{x}_i$, and $\mathbf{u}_i$ represents sampling noise. In general, $d \ll D$, that is, the dimension of the manifold is much smaller than that of the input space. It is assumed that there is sufficient data such that the manifold is well-sampled.

One important geometric intuition behind the LLE algorithm is that each data point and its neighbors lie on or are close to a locally linear patch of the manifold. LLE tries to characterize the geometry of the local patches by finding the linear coefficients that reconstruct each data point from its neighbors. Let $\mathcal{N}_i$ be the set of k nearest neighbors of $\mathbf{x}_i$ (not including $\mathbf{x}_i$ itself). Then the reconstruction coefficient can be computed by minimizing the reconstruction error, which is measured as

$$\Gamma(W) = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{\mathbf{x}_j \in \mathcal{N}_i} w_{ij}\mathbf{x}_j\|_2^2. \tag{2}$$

The reconstruction error is minimized subject to the constraint that the rows of the weight matrix sum to one: $\sum_j w_{ij} = 1$.

Let $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$. Then Y can be computed by minimizing the embedding cost function

$$\Phi(Y) = \sum_{i=1}^{n} \|\mathbf{y}_i - \sum_j w_{ij}\mathbf{y}_j\|_2^2 = YMY^T, \tag{3}$$

where M is given by

$$M_{ij} = \delta_{ij} - w_{ij} - w_{ji} + \sum_k w_{ki}w_{kj}. \tag{4}$$

Here δ_{ij} is 1 if $i = j$ and 0 otherwise. The mapping cost, $\Phi(Y)$, is translation and rotation invariant. To make the problem well-posed, the cost function is minimized subject to the constraints that $\sum_i \mathbf{y}_i = \mathbf{0}$, and that $\sum_i \mathbf{y}_i\mathbf{y}_i^T = I$, where I is the identity matrix. The resulting problem is equivalent to finding the smallest $d + 1$ eigenvectors of the matrix M.

The LTSA algorithm tries to characterize the local geometry by computing an approximate tangent space at each data point. Let the Jacobian matrix of $\mathbf{f}$ at $\mathbf{y}$ be

$$J_{\mathbf{f}}(\mathbf{y}) = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_d} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_D}{\partial y_1} & \cdots & \frac{\partial f_D}{\partial y_d} \end{bmatrix}. \tag{5}$$

The tangent space $\mathcal{T}_{\mathbf{y}}$ of $\mathbf{f}$ at $\mathbf{y}$ is defined as the subspace spanned by the d columns of the $J_{\mathbf{f}}(\mathbf{y})$, that is, $\mathcal{T}_{\mathbf{y}} = span(J_{\mathbf{f}}(\mathbf{y}))$. Without knowing the function $\mathbf{f}$ we cannot compute the Jacobian matrix $J_{\mathbf{f}}(\mathbf{y})$. However, $\mathcal{T}_{\mathbf{y}}$ can be approximated by the subspace spanned by the first d principal components of a neighbor set of $\mathbf{y}$.

Once the tangent space at each data point has been computed, the global coordinates are computed by "aligning" the local tangent spaces together. Let $\mathcal{N}_i$ be the set of k nearest neighbors of $\mathbf{y}_i$ (including $\mathbf{y}_i$ itself). Denote the neighborhood index set of $\mathbf{y}_i$ as $\mathcal{I}_i$. Let $\mathbf{g}_{i1}$, $\mathbf{g}_{i2}$, ..., $\mathbf{g}_{id}$ be the d principal components of $\mathcal{N}_i$. Let $G_i = [\mathbf{e}/\sqrt{k}, \mathbf{g}_{i1}, \mathbf{g}_{i2}, \ldots, \mathbf{g}_{id}]$. It was shown that the global coordinates can be computed by minimizing the alignment cost

$$\Phi(Y) = YMY^T. \tag{6}$$

Here M is the alignment matrix computed as follows:

$$M(\mathcal{I}_i, \mathcal{I}_i) \leftarrow M(\mathcal{I}_i, \mathcal{I}_i) + I - G_iG_i^T, \quad i = 1, 2, \ldots, n \tag{7}$$

with M initially set to 0. It was shown (Zha & Zhang, 2005) that under certain conditions, M has $d+1$ zero eigenvalues, and that the null space of M spans the low dimensional coordinate space. As in LLE, the cost function is translation and rotation invariant. Therefore, it is minimized subject to the constraints $\sum_i \mathbf{y}_i = \mathbf{0}$ and $\sum_i \mathbf{y}_i \mathbf{y}_i^T = I$, and the resulting problem can be solved by computing the $d+1$ smallest eigenvectors of M.

ISOMAP (Tenebaum et al., 2000) is based on the classical MDS, but seeks an embedding that preserves the pairwise geodesic manifold distance rather than the Euclidean distance. The geodesic distances are approximated by "adding up a sequence of short hops between neighboring points", which are computed by finding the shortest paths in a graph with edges connecting only neighboring data points. Let Δ be the matrix of squared geodesic distances. Let P be the $n \times n$ projection matrix $I - \mathbf{e}\mathbf{e}^T/n$, where $\mathbf{e} = [1, \ldots, 1]^T \in \mathcal{R}^n$. Then the low dimensional global coordinates are computed by finding the d maximum eigenvectors of

$$A = -\frac{1}{2} P^T \Delta P, \tag{8}$$

each scaled by the square root of its corresponding eigenvalue.

3. DERIVATION OF THE SS-LLE, SS-LTSA, AND SS-ISOMAP ALGORITHMS

The SS-LLE algorithm inherits the basic idea of LLE, that is, it tries to characterize the local geometry by the reconstruction weights, and finds the global low dimensional coordinates by minimizing the embedding cost. In the presence of prior information, the reconstruction weights can be computed the same way as was done in the basic LLE algorithm, but the embedding cost function is minimized subject to the constraint that the low dimensional coordinates obey prior information. Similarly, the SS-LTSA algorithm captures the local geometry by computing an approximate tangent space the same way as the basic LTSA algorithm, but computes an alignment that obeys prior information.

Suppose the exact mapping of m data points is known. Note that if $m \leq d+1$, then in both the LLE and LTSA algorithm, the prior information only helps to remove the freedom of translation and scaling. For the rest of this paper, unless otherwise specified, it is assumed that $m > d+1$. Without loss of generality, assume that it is the first m data points whose low dimensional coordinates are known. Partition Y as $[Y_1 \; Y_2]$, where Y_1 corresponds to the data points whose low dimensional coordinates are known, and Y_2 corresponds to the other data points. Partition M as follows:

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix}, \tag{9}$$

where M_{11} is a matrix of size $m \times m$. For both the SS-LLE and SS-LTSA, since Y_1 is known, the minimization problem can be written as

$$\min_{Y_2} [Y_1 \; Y_2] \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix} \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix}, \tag{10}$$

or equivalently

$$\min_{Y_2} Y_2 M_{22} Y_2^T + 2 Y_1 M_{12} Y_2^T. \tag{11}$$

By setting the gradient of the above objective function to zero, we get

$$M_{22} Y_2^T = M_{12} Y_1^T. \tag{12}$$

We see that the global low dimensional coordinates can be computed by solving a linear system of equations.

In order to derive the SS-ISOMAP algorithm, we first restate the basic ISOMAP problem as follows:

$$\max_Y Y A Y^T \quad \text{subject to } Y Y^T = I, \tag{13}$$

where A is the matrix given in (8). Let $A = Q \Lambda Q^T$ be the eigen-decomposition of A. Let $Q = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$, let $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Define the matrix M as follows:

$$M = \lambda_1 I - A - \sum_{i=2}^{d} (\lambda_1 - \lambda_i) \mathbf{q}_i \mathbf{q}_i^T - \lambda_1 \mathbf{e}\mathbf{e}^T/n. \tag{14}$$

Then it is easy to check that M has $d+1$ zero eigenvalues, furthermore, its null space is given by $span([\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_d, \mathbf{e}])$. Therefore (13) can be equivalently stated as

$$\min_Y Y M Y^T \quad \text{subject to } Y Y^T = I, \sum_{i=0}^{n} \mathbf{y}_i = 0. \tag{15}$$

It follows that the SS-ISOMAP solution can be obtained by solving a minimization problem that has the same form as (10), but M is replaced with the matrix given in (14). This can be solved much the same way as SS-LLE and SS-LTSA. However, experimental results indicate that other than being able to map input data to a properly scaled and translated space, the improvement of SS-ISOMAP over the basic ISOMAP is not significant.

4. SENSITIVITY ANALYSIS, CHOOSING PRIOR POINTS

If Y_1 is exact, then the error in (12) only exists in M_{22} and M_{12}. Consider the parameterized system:

$$(M_{22} + \epsilon F)Y_2(\epsilon)^T = (M_{12} + \epsilon G)Y_1^T. \tag{16}$$

According to a result from (Golub & Van Loan, 1996),

$$\frac{\|Y_2(\epsilon) - Y_2\|_2}{\|Y_2\|_2} \leq \kappa(M_{22})|\epsilon|(\rho_A + \rho_B), \tag{17}$$

where $\kappa(M_{22})$ denotes the condition number of M_{22}, $\rho_A = \|F\|/\|M_{22}\|$ and $\rho_B = \|F\|/\|M_{12}\|$ for some appropriate matrix norm. That is, relative error in the computed low dimensional coordinates can be bounded by $\kappa(M_{22})$ times the relative error in M_{22} and M_{12}. In other words, the sensitivity of the computed solution depends on the condition number of the matrix M_{22}.

According to the interlacing property theorem (Golub & Van Loan, 1996, §8.1.2), if $M \in \mathcal{R}^{n\times n}$ is symmetric and $M_r = M(1:r, 1:r)$, then

$$\lambda_{r+1}(M_{r+1}) \leq \lambda_r(M_r), \text{ and } \lambda_1(M_r) \leq \lambda_1(M_{r+1}). \tag{18}$$

Here $\lambda_i(M_j)$ denotes the i-th eigenvalue of M_j. It follows that if $\lambda_{r+1}(M_{r+1})$ is positive, then the condition number of M_r, which equals $\lambda_1(M_r)/\lambda_r(M_r)$, is smaller than that of M_{r+1}. Therefore, with the increase of prior points, the condition number of the coefficient matrix in (12) gets smaller and smaller, and the computed solution gets less and less sensitive to noise in M_{22} and M_{12}.

Because the sensitivity of the solution depends on the condition number of the matrix M_{22}, the most helpful prior points are those that help reduce the condition number of M_{22} the most. Consider the simplest case, in which we have a one dimensional manifold, and all the sample data points are evenly spaced. Without loss of generality, assume that the samples are ordered as they appear on the manifold. Then the resulting mapping matrix in LLE or alignment matrix in LTSA will be banded and exhibit a Toeplitz structure (Chan & Ng, 1996) except at the first and last few rows/columns. For example, if $k = 3$, then the alignment matrix in the LTSA algorithm is as follows:

$$M = \frac{1}{6}\begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix} \tag{19}$$

For this particular case, assuming that the first two prior points are the end points, it can be shown that if n is sufficiently large, the condition number of M_{22} is the smallest if additional prior points are chosen such that the minimum pairwise distance between the prior points is maximized. Table 1 shows the condition number of M_{22} when $n = 500$ and if the above mentioned strategy is used to select prior points. As can be seen, the condition number of M_{22} can be considerably reduced with just a few prior points.

Table 1. $\kappa(M_{22})$ when different number of prior points are used, $n = 500$.

No. of prior points	$\kappa(M_{22})$
2	1.9977e+09
3	2.6291e+08
5	2.9461e+07
9	2.2904e+06
17	1.5307e+05

For most real life problems, $d > 1$, furthermore, the data points are sampled irregularly, thus, analyzing the condition number of the mapping or alignment matrix will be much more complicated. However, the idea of maximizing minimum pairwise geodesic distance between the prior points can be generalized to real life problems. Unfortunately, finding such a combination of prior points is still a very difficult task. Fortunately, a near optimal solution can be computed by using a technique that resembles the one Silva and Tenenbaum (de Silva & Tenenbaum, 2004) used to select landmark points for sparse MDS: Prior points are chosen one at a time, and each new prior point maximizes, over all unused data points, the minimum geodesic distance to any of the existing prior points. The first prior point is chosen arbitrarily. Note that this procedure needs no information on the underlying low dimensional coordinates.

The analysis of the condition number of the matrix given in (14) is quite different. Indeed, SS-ISOMAP is not as effective as SS-LLE and SS-LTSA.

5. INEXACT PRIOR INFORMATION

For inexact prior information, we propose to minimize an objective function that combines the mapping error with a regularization term, weighted by a parameter. Thus

$$\Phi(Y_1, Y_2) = [Y_1\ Y_2]\begin{bmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix}\begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix} + \beta\|Y_1 - \hat{Y}_1\|_F^2. \tag{20}$$

Here $\hat{Y}_1$ represents prior information, $\|\cdot\|_F$ denotes Frobenius norm, and β is the regularization parameter that reflects our confidence level in prior information. If we are fully confident in the provided prior information, then $\beta \to \infty$, and the resulting problem is equivalent to (12). If the prior information is totally not trustworthy, then $\beta = 0$, and the problem is equivalent to an unsupervised problem.

The objective function of (20) is quadratic. Under weak assumptions, it can be shown that this function has a symmetric positive definite Hessian matrix, therefore, its minimizer can be computed by solving the following linear system of equations:

$$\begin{bmatrix} M_{11}+\beta I & M_{12} \\ M_{12}^T & M_{22} \end{bmatrix} \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix} = \begin{bmatrix} \beta \hat{Y}_1^{\,T} \\ 0 \end{bmatrix} \tag{21}$$

6. EXPERIMENT RESULTS

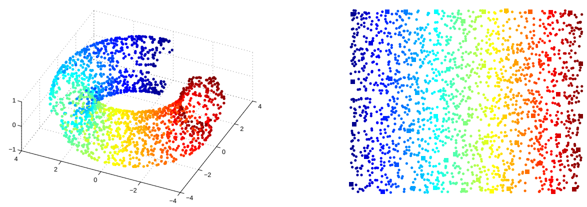
First, we apply both the semi-supervised and unsupervised algorithms to the data set sampled from the "incomplete tire" (a tire with a slice and a strip cut out) shaped manifold shown in Figure 1(a). The data points are generated by the following MATLAB commands:

```
t = pi*5*rand(1,N)/3;
s = pi*5*rand(1,N)/3;
X = [(3+cos(s)).*cos(t);
     (3+cos(s)).*sin(t);
     sin(s)];
```

A total of $n = 2000$ data points are sampled. The generating low dimensional coordinates are shown in Figure 1(b).

Note that the "incomplete tire" is quite different from the swiss roll used in (Roweis & Saul, 2000), (Tenebaum et al., 2000) and the S-curve used in (Saul & Roweis, 2003), because the swiss roll and S-curve can both easily be "flattened out". In other words, there exists an isometric mapping that maps the swiss roll or S-curve to a two dimensional linear space, but there is no mapping that maps the incomplete tire to a linear space while preserving all manifold distances. Consequently, the "incomplete tire" poses more challenges to dimensionality reduction algorithms than the swiss roll and S-curve. We remark that the "incomplete tire" is a better example of real life manifolds, since it is very unlikely that real life manifolds can be mapped to a low dimensional linear space by an isometric mapping.

Figure 2 shows the two dimensional embedding computed by the basic and semi-supervised algorithms using 50 prior points. It can be seen that the basic algorithms handle the challenges posed by the "incomplete tire" poorly, but SS-LLE and SS-LTSA yield remarkably good results. Furthermore, the semi-supervised algorithms are less sensitive to the number of neighbors.



(a) (b)

Figure 1. (a) the "incomplete tire", (b) the generating coordinates

Figure 3 shows the relative error of the semi-supervised solutions as compared to the true underlying low dimensional coordinates when the number and locations of the prior points vary. These results confirm the theoretical prediction that increasing the number of prior points decreases the relative error of the solution, and that carefully chosen prior points better improves the solution than randomly spaced prior points.

In the following example, we use the dataset from (Rahimi et al., 2005), which shows a subject moving his arms, and was used for upper body tracking. We choose 2000 frames from this video sequence. The frames are downsampled by the pixel size such that Matlab can load the data into the main memory. Figure (5) shows 20 frames, which we use as prior points. And the locations of the elbows and wrists, which are marked in blue in the figures, are are manually determined by a human. Here we apply our semisupervised algorithms to find the locations of the elbows and wrists in other frames. Considering the location of each elbow/wrist has two dimensions, we preset the dimensionality of the manifold to be 8.

Figure (4) shows the elbow and wrist locations of certain frames recovered by out algorithm SSLTSA with 24 nearest neighbors. As can be seen, they coincide with the real locations very well. In order to test our algorithm for inexact prior information that was presented in section 5, we artificially added a 5 % noise to wrist and elbow locations of the prior frames, and applyed the algorithm in section 5 with $\beta = 10$. The results are shown in Figure (6), compared with Figure (4), it can be seen that our algorithm returns good results even when the prior information is ineact. The optimal regularization parameter β can be chosen ei-

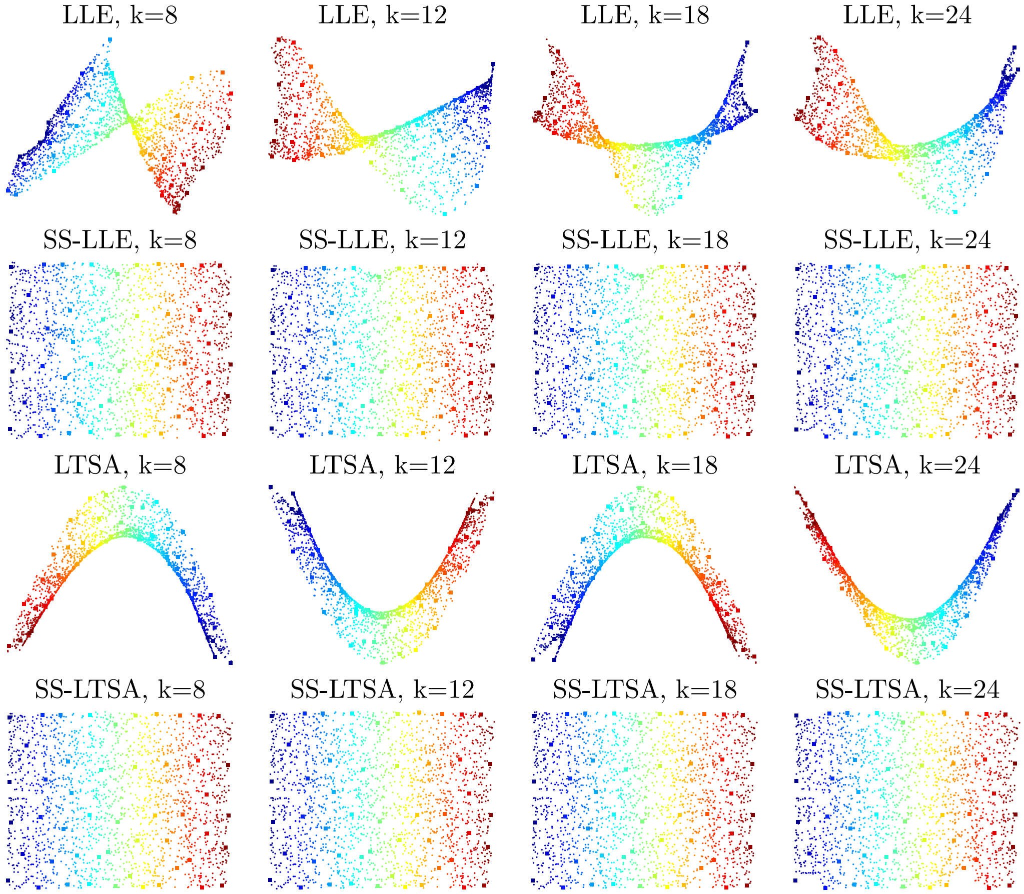


Figure 2. Two dimensional coordinates computed by the basic and semi-supervised algorithms, using different number of neighbors.

ther by the L-curve method or Cross Validation. In fact, our experimental results indicate that there is no need for such sophisticated schemes, since the results are quite good for a very wide range of β values.

7. CONCLUSIONS

In conclusion, we have proposed semi-supervised algorithms for nonlinear dimensionality reduction. These algorithms compute a low dimensional embedding that minimizes mapping cost subject to the condition that the low dimensional coordinates obey prior information. Theoretical analysis and experimental results indicate that prior information helps improve the solution.

8. ACKNOWLEDGMENTS

This work was supported in part by NSF grants CCF-0305879, CCF-0429481 and DMS-0311800. ALso we would like to thank Dr. Ali Rahimi, who shared their upper body tracking data with us.

References

Chan, R., & Ng, M. (1996). Conjugate gradient methods for Toeplitz systems. *SIAM Review*, *38(3)*, 427–482.

de Silva, V., & Tenenbaum, J. (2004). *Sparse multidimensional scaling using landmark points* (Technical Report). Stanford University.

Golub, G., & Van Loan, C. (Eds.). (1996). *Matrix computations*. Baltimore: The Johns Hopkins University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (Eds.). (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Rahimi, A., Recht, B., & Darrell, T. (2005). Learning

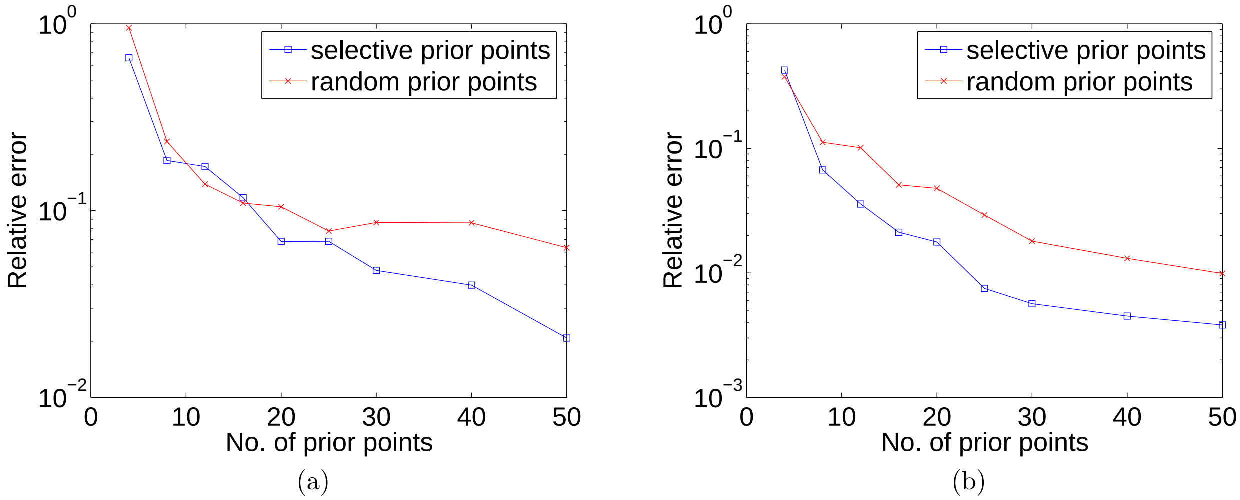


Figure 3. (a) relative error of the solutions computed by SS-LLE, (b) relative error of the solutions computed by SS-LTSA.

Figure 4. 10 frames from the results of SSLTSA .

appearance manifolds from video. *Computer Vision and Pattern Recognition (CVPR)*.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290(5500)*, 2323–2326.

Saul, L., & Roweis, S. (2003). Think globally, fit locally: unsupervised learning of nonlinear manifolds. *Journal of Machine Learning Research, 4*, 119–155.

Tenebaum, J., de Silva, V., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290(5500)*, 2319–2323.

Weinberger, K., Packer, B., & Saul, L. (2005). Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. *Proceedings of the tenth international workshop on artificial intelligence and statistics*, 381–388.

Zha, H., & Zhang, Z. (2005). Spectral analysis of alignment in manifold learning. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Zhang, Z., & Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing, 26(1)*, 313–338.

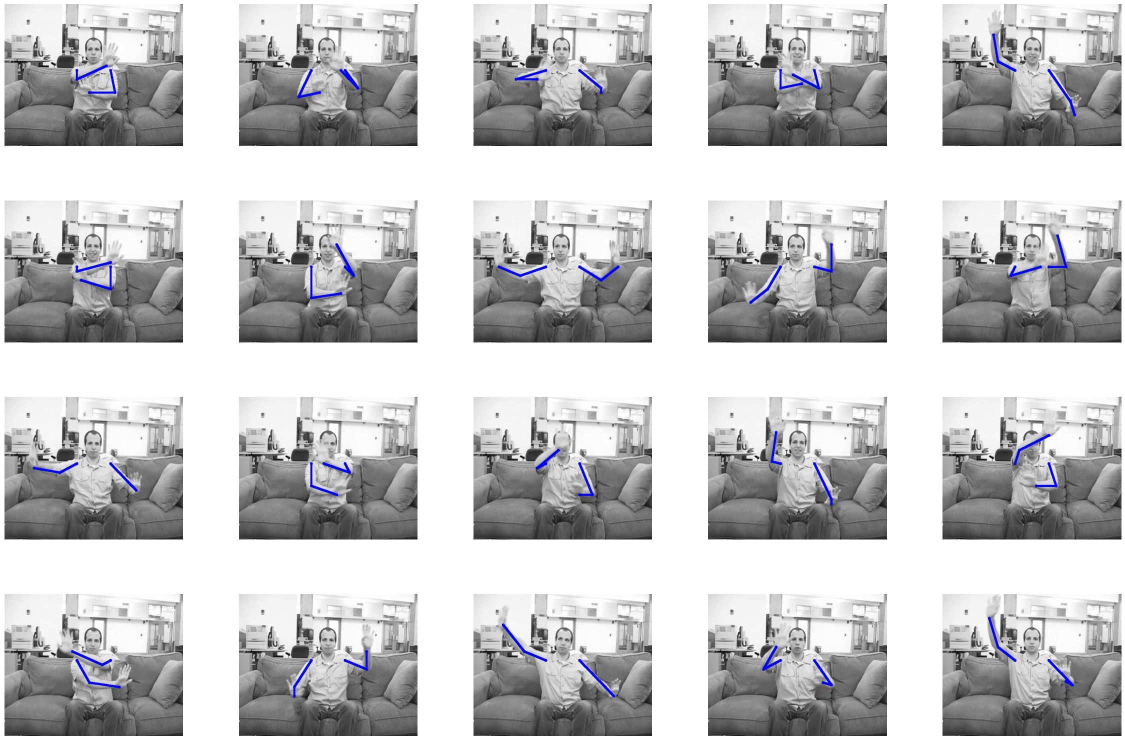


Figure 5. 20 frames with prior information, which are the locations of the elbows/wrists.

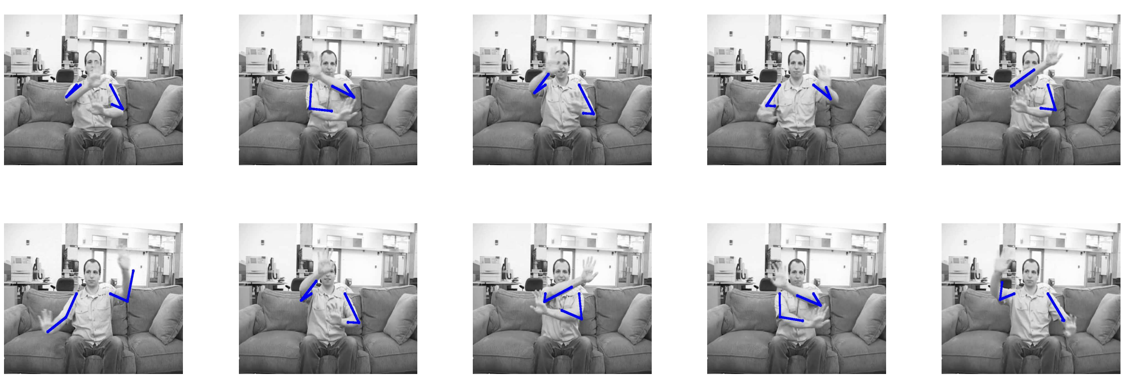


Figure 6. The results of inexact prior informations algorithm.