ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2007, Vol. 36, No. 4

SELECTION OF THE NUMBER OF NEIGHBOURS OF EACH DATA POINT FOR THE LOCALLY LINEAR EMBEDDING ALGORITHM

Rasa Karbauskaitė[1,2], Olga Kurasova[1,2], Gintautas Dzemyda[1,2]

[1] *Institute of Mathematics and Informatics,*
Akademijos St. 4, 08663, Vilnius, Lithuania
[2] *Vilnius Pedagogical University*
Studentų St. 39, 08106, Vilnius, Lithuania

Abstract. This paper deals with a method, called locally linear embedding. It is a nonlinear dimensionality reduction technique that computes low-dimensional, neighbourhood preserving embeddings of high dimensional data and attempts to discover nonlinear structure in high dimensional data. The implementation of the algorithm is fairly straightforward, as the algorithm has only two control parameters: the number of neighbours of each data point and the regularisation parameter. The mapping quality is quite sensitive to these parameters. In this paper, we propose a new way for selecting the number of the nearest neighbours of each data point. Our approach is experimentally verified on two data sets: artificial data and real world pictures.

Keywords: locally linear embedding; dimensionality reduction; manifold learning.

1. Introduction

Data coming from the real world are often difficult to understand because of their high dimensionality. A number of dimensionality reduction techniques are proposed, that allow the user to better analyse or visualize complex data sets.

Dimensionality reduction techniques may be divided into two classes. In the first one, there are linear methods, such as the Principal Component Analysis (PCA, [7]), or the classical scaling ([4, 5]), etc. However, the underlying structure of real data is often highly nonlinear and hence cannot be approximated by linear manifolds. The second class includes nonlinear algorithms, such as nonlinear variants of multidimensional scaling (MDS) [4, 5], the self-organising map (SOM) [8], generative topographic mapping (GTM) [3], principal curves and surfaces [6], etc.

Several nonlinear manifold learning methods – locally linear embedding (LLE) [11, 12], Isomap [13], Laplacian Eigenmaps [2] – have been developed recently. These methods are supposed to overcome the difficulties experienced with other classical nonlinear approaches mentioned above: they are simple to implement, have a very small number of free parameters, and do not trap local minima. These algorithms are able to recover the intrinsic geometric structure of a broad class of nonlinear data manifolds and come in two flavours: local and global. Local approaches (e.g., LLE, Laplacian Eigenmaps) attempt to preserve the local geometry of the data; particularly, they seek to map nearby points on the manifold to nearby points in the low-dimensional representation. Global approaches (e.g., Isomap) attempt to preserve geometry at all scales, by mapping nearby points on the manifold to nearby points in a low-dimensional space, and faraway points to faraway points.

In this paper, we concentrate on the LLE algorithm. What are the advantages of LLE compared with PCA and MDS? The dimensionality reduction by LLE succeeds in identifying the underlying structure of the manifold, while PCA or MDS methods map faraway data points on the manifold to nearby points in the plane, failing to identify the structure. Unlike MDS, LLE eliminates the need to estimate pairwise distances between widely separated data points. This fact is illustrated in Figure 2, by mapping a nonlinear two-dimensional S-manifold (Figure 1).

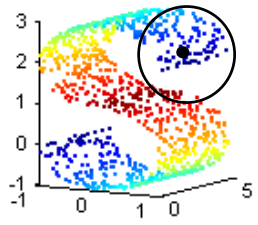
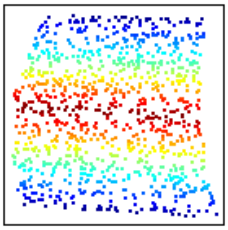
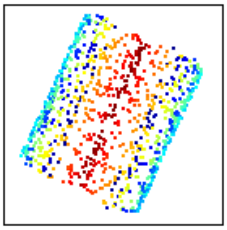


Figure 1. A nonlinear S-manifold consisting of 1000 data points


a) the result obtained by LLE


b) the result obtained by PCA

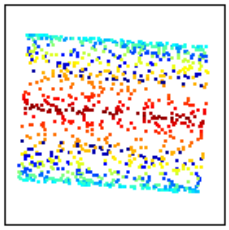

c) the result obtained by MDS

Figure 2. Embeddings of the S-manifold, obtained by different methods

The main control parameter of the LLE algorithm is the number of neighbours of each data point. This parameter strongly influences the results obtained. We propose here a new way for selecting the number of the nearest neighbours of each data point and apply LLE to high dimensional data visualization.

2. Locally linear embedding method

Locally linear embedding (LLE) [11, 12] is a non-linear method for dimensionality reduction and manifold learning. Given a set of data points distributed on a manifold in a high dimensional space, LLE is able to project the data to a lower space by unfolding the manifold.

LLE works by assuming that the manifold is well sampled, i.e., there are enough data, each data point and its neighbours lie on or close to a locally linear patch. Therefore, a data point can be approximated as a weighted linear combination of its neighbours. The basic idea of LLE is that such a linear combination is invariant under linear transformations (translation, rotation, and scaling) and, therefore, should remain unchanged after the manifold has been unfolded to a low space. The low dimensional configuration of data points is given by solving two constrained least squares optimisation problems.

The input of the LLE algorithm consists of m $n-$ dimensional vectors $\boldsymbol{X}_i, i=1,...,m$ ($\boldsymbol{X}_i \in R^n$). The output consists of m $d-$ dimensional vectors $\boldsymbol{Y}_i, i=1,...,m$ ($\boldsymbol{Y}_i \in R^d$). The LLE algorithm has three steps. In the first step, one identifies k neighbours of each data point $\boldsymbol{X}_i$. Different criteria for neighbour selection can be adopted; the simplest possibility is to choose the k-nearest neighbours according to the Euclidean distance. In the second step, one computes the weights w_j^i that reconstruct each data point $\boldsymbol{X}_i$ best from its neighbours $\boldsymbol{X}_{N(1)},...,\boldsymbol{X}_{N(k)}$, minimizing the following error function

$$E(W)=\sum_{i=1}^{m}\left|\boldsymbol{X}_i-\sum_{j=1}^{k}w_j^i\boldsymbol{X}_{N(j)}\right|^2,$$

subject to the constraints $\sum_{j=1}^{k}w_j^i=1$ and $w_j^i=0$, if $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are not neighbours. This is a typical constrained least squares optimisation problem, which can be easily answered by solving a linear system of equations. The third step consists in mapping each data point $\boldsymbol{X}_i$ to a low-dimensional vector $\boldsymbol{Y}_i$, which best preserve high-dimensional neighbourhood geometry represented by the weights w_j^i. That is, the weights are fixed and we need to minimize the following function:

$$\Phi(Y)=\sum_{i=1}^{m}\left|\boldsymbol{Y}_i-\sum_{j=1}^{k}w_j^i\boldsymbol{Y}_{N(j)}\right|^2,$$

subject to two constraints: $\sum_{i=1}^{m}\boldsymbol{Y}_i=0$ and $\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{Y}_i\boldsymbol{Y}_i^T=\boldsymbol{I}$, where $\boldsymbol{I}$ is the $d\times d$ identity matrix, those provide a unique solution. The most straightforward method for computing the d-dimensional coordinates ($d<n$) is to find the bottom $d+1$ eigenvectors of the sparse matrix $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^T(\boldsymbol{I}-\boldsymbol{W})$, ($\boldsymbol{W}=(w_j^1,w_j^2,...,w_j^m)$, $j=1,...,k$). These eigenvectors are associated with the $d+1$ smallest eigenvalues of $\boldsymbol{M}$. The bottom eigenvector, whose eigenvalue is closest to zero, is the unit vector with all equal components and it is discarded. The remaining d eigenvectors form the d embedding coordinates that are found by LLE.

3. Selection of the number of the nearest neighbours

The most important step to success of LLE is the first step, that is, to define the number k of the nearest neighbours for each data point. The mapping quality is rather sensitive to this parameter. If k is set too small, the continuous manifold can falsely be divided into disjoint sub-manifolds, in this way, the mapping does not reflect any global properties (Figure 3, for example $k=5$). If k is too high, a large

number of the nearest neighbours causes smoothing or elimination of small-scale structures in the manifold, the mapping loses its nonlinear character (Figure 3, for example $k = 100$) and behaves like traditional PCA (Figure 2b).

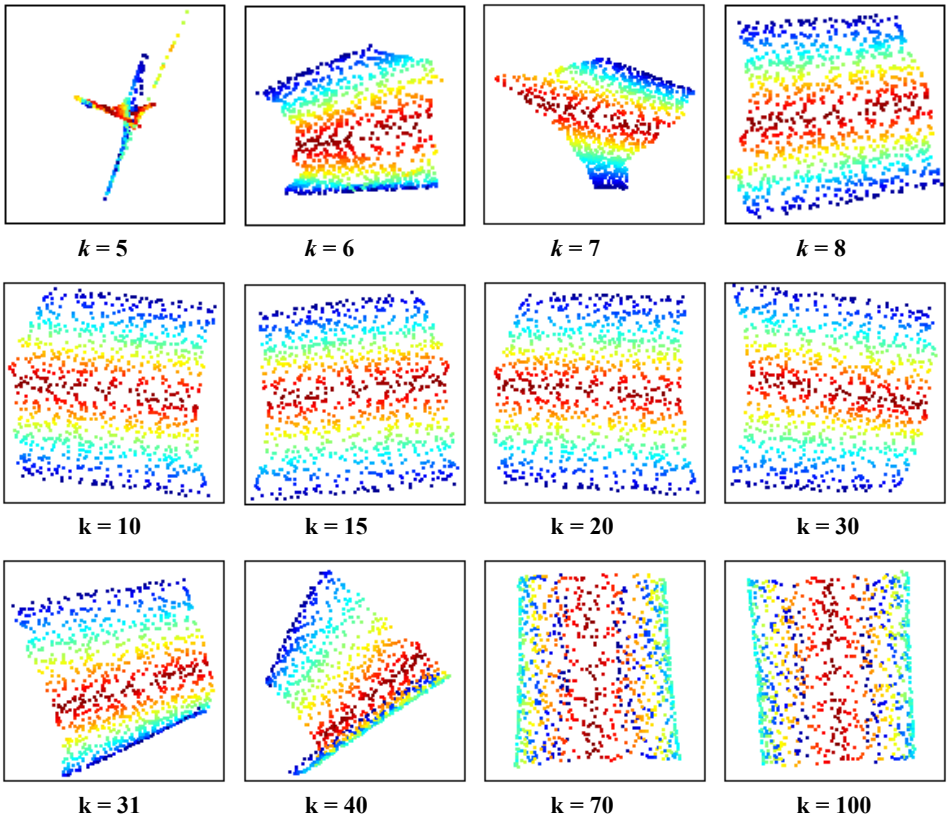
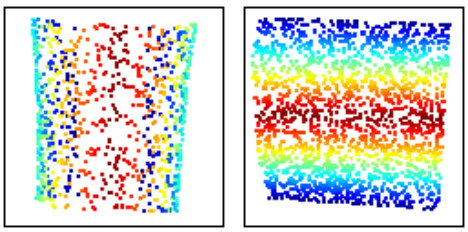


Figure 3. Embeddings of the 2-dimensional S-manifold, computed for different choices of the number of the nearest neighbours k by LLE



a) $m = 1000$, $k = 50$ b) $m = 2000$, $k = 50$

Figure 4. Embeddings of the S-manifolds with LLE

The results of LLE [12] are typically stable over some range of neighbourhood sizes. Figure 3 shows a range of embeddings discovered by the LLE algorithm, all on the same data set, but using different numbers of the nearest neighbours k. A reliable embedding is obtained over a wide range of values, i.e., $k \in [8; 30]$. However, as mentioned in [12], the size of that range depends on various features of the data, such as the sampling density and manifold geometry. The dependence of LLE results on sampling density is shown in Figure 4. Two 2-dimensional S-manifolds were investigated. One of them consisted of 1000 points and the other of 2000 points. In both cases, embeddings were computed, as $k = 50$. LLE failed to unravel the S-manifold of 1000 points and succeeded in unraveling the manifold of 2000 points.

If the structure of the manifold is known in advance, we can use a subjective evaluation that accompanies a human visual check. But what can we say about the reliability of the embeddings computed using a certain value of the parameter k, when the structure of the manifold is not clear? To estimate the embeddings, it is necessary to use quantitative numerical measures. Spearman's rho or the residual variance is commonly used for estimating the topology preservation with a view to reduce dimensionality. Automatic selection of the number of the nearest neighbours was proposed in [9].

3.1. A new way for selecting a proper range of neighbourhood sizes

As shown in Figure 3, it is not necessary to find the optimal number of the nearest neighbours, but it is enough to estimate a proper range of neighbourhood sizes. In this paper, we propose a new way for solving this problem. In order to quantitatively estimate the topology preservation, we compute Spearman's rho. It estimates the correlation of rank order data, i.e., how well the corresponding low-dimensional projection preserves the order of the pairwise distances between the high-dimensional data points converted to ranks. Spearman's rho is computed by using the following equation:

$$\rho_{Sp} = 1 - \frac{6\sum_{i=1}^{T}\left(r_x(i) - r_y(i)\right)^2}{T^3 - T},$$

where T is the number of distances to be compared, $r_x(i)$ and $r_y(i)$ are the ranks of the pairwise distances calculated for the original and projected data points. $-1 \le \rho_{Sp} \le 1$. The best value of Spearman's rho is equal to one.

In the calculation of Spearman's rho, distances both on the plane and on a multidimensional space are used. A question arises which distances should be evaluated when estimating Spearman's rho: Euclidean or geodesic? Euclidean distances are usually used on the plane. On a multidimensional space, either the Euclidean or geodesic distances are applied. Geodesic distances represent the shortest paths along the curved surface of the manifold. The author in [1] states that the Euclidean distance is not good for finding the shortest path between points within the framework of the manifold. The paper [13] states that it is necessary to apply geodesic distances in order to preserve the global structure of the manifold. It is reasonable to use the Euclidean distances in case the manifold is flat, therefore in further experiments on the plane we will always evaluate only Euclidean distances.

The S-manifold ($m = 1000$) has been investigated. The LLE algorithm was run for many times gradually increasing the number of neighbours $k \in [5; 100]$, each time calculating Spearman's rho (Figure 5). Two dependences of Spearman's rho on k have been obtained: (I) the Euclidean distances were evaluated in a space, (II) the geodesic distances were evaluated in a space. Let the number of neighbours be $k = 100$. We see that, when estimating the Euclidean distances, the value of Spearman's rho is near to 1 (≈ 0.97), and when estimating the geodesic distances in a space, the value of Spearman's rho is much lower (≈ 0.82). If $k = 100$, the Euclidean distances are preserved very well, but the structure of the manifold is destroyed (Figure 3, $k = 100$), and we wished to preserve it. This experiment corroborates the fact that it is indispensable to evaluate geodesic distances in a space; therefore we will evaluate only geodesic distances in our further experiments.

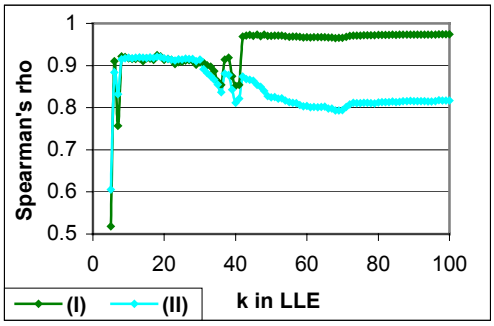


Figure 5. Dependences of Spearman's rho on k obtained after visualizing the S-manifold by LLE: (I) Euclidean distances were evaluated in a space, (II) geodesic distances were evaluated in a space

Only one parameter is selected in the calculation algorithm of geodesic distances – the number of the nearest neighbours necessary to draw a graph. Denote it as k_{geod}. The LLE algorithm also has the same kind of parameter, – the number of neighbours k. What value of k_{geod} should it be when calculating geodesic distances? Should k_{geod} be coincident with the chosen number of neighbours k in the LLE algorithm?

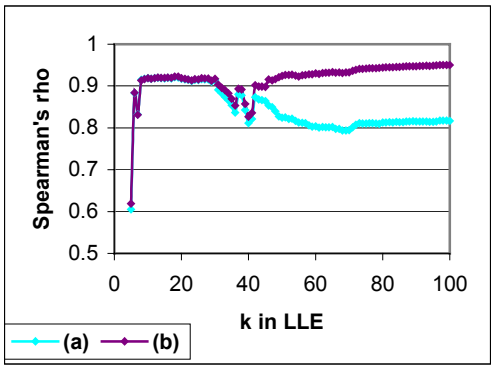


Figure 6. Dependences of Spearman's rho on k obtained after visualizing the S-manifold by LLE. Geodesic distances were evaluated in a space, as (a) $k_{geod} = 10$, (b) $k_{geod} = k$

In Figure 6, two dependences of Spearman's rho on k have been obtained: (a) when calculating geodesic distances in a space, a very small number of neighbours was fixed, e.g., $k_{geod} = 10$, (b) when calculating geodesic distances, the number of neighbours was varying just like in the LLE algorithm, i.e., $k_{geod} = k$. If $k = 100$, the value of Spearman's rho according to curve (a) is rather low (≈ 0.82), and the declined curve rises but slightly. Hence it follows that distances are badly retained and the mapping does not

represent the global structure. Curve (b) illustrates that the value of Spearman's rho approaches 1 (≈ 0.95). It implies that the LLE result is rather good. However it is obvious that after visualising these data by LLE with $k = 100$, the resulting mapping does not reflect the structure of the manifold (Figure 3, $k = 100$), though the value of Spearman's rho is close to 1. The reason why is as follows: if very many neighbours k_{geod} are selected while calculating geodesic distances in a space, then the structure of nonlinear manifold is destroyed, i.e., the nearest neighbours to a point in a space may be the points met in the transition across the manifold (Euclidean distances are calculated when looking for neighbours). In Figure 1, the neighbours of the point marked by a black circular disk fall into the black circle. In this case, the LLE algorithm contains as many neighbours as that for calculating geodesic distances: $k = k_{geod}$ (neighbours in the LLE algorithm are found by calculating Euclidean distances). Therefore, faraway points on the manifold are treated as the close ones both in a space and in a plane. This is the reason why the value of Spearman's rho increases with an increase in number of the nearest neighbours. Good embeddings in Figure 3 are obtained when curve (a) in Figure 6 reaches its maximum. Therefore, Spearman's rho with fixed rather small k_{geod} may be used as criterion for visualization quality.

4. Application of LLE in analysis of picture set

One of the applications of the LLE method in practice is visualization of the points, the coordinates of which are comprised of the parameters of pictures. A picture is digitised, i.e., a vector consists of colour parameters of pixels therefore it is of very large dimension. The particularity of these data is that, the data are comprised of pictures of the same object, by turning the object gradually at a certain angle. In this way the points differ from one another slightly, making up a certain manifold. For an experiment uncoloured pictures were used, obtained by gradually rotating a duckling at the 360° angle [10]. The number of pictures (points) was $m = 72$. The images had 128×128 grayscale pixels, therefore the dimension of points in a multidimensional space is $n = 16384$. The LLE algorithm was run for 35 times as $k \in [2; 36]$. Each time Spearman's rho was calculated. Three dependences of Spearman's rho on k have been shown in Figure 7: (I), when calculating geodesic distances in a space, a very small number of neighbours was fixed, e.g., $k_{geod} = 2$; (II), when calculating geodesic distances, the number of neighbours is varying just like in the LLE algorithm, i.e., $k_{geod} = k$; (III) – Euclidean distances were estimated in a space. We see that cases (I) and (II) bear the highest values of Spearman's rho, i.e., $0.91 \le \rho_{Sp} \le 0.97$ as $k \in [2; 8]$, while case (III) has much lower values of Spearman's rho as $k \in [2; 8]$: $0.66 \le \rho_{Sp} \le 0.7$. For $k \ge 9$, the values of Spearman's rho considerably diminish ($\rho_{Sp} \approx 0.54$, as $k = 9$) in case (I), in case (II) they decrease a little less ($\rho_{Sp} \approx 0.82$, as $k = 9$), and in case (III), on the contrary, the values increase. Embeddings, obtained after visualising these data by LLE, are presented in Figure 8. Since the object was gradually turned round at the 360° angle, it is likely that the true representation is obtained in Figure 8a as $k \in [2; 8]$. Hence it follows that cases (I) and (II) yield the right result. Case (I) illustrates an especially explicit difference between these solutions.

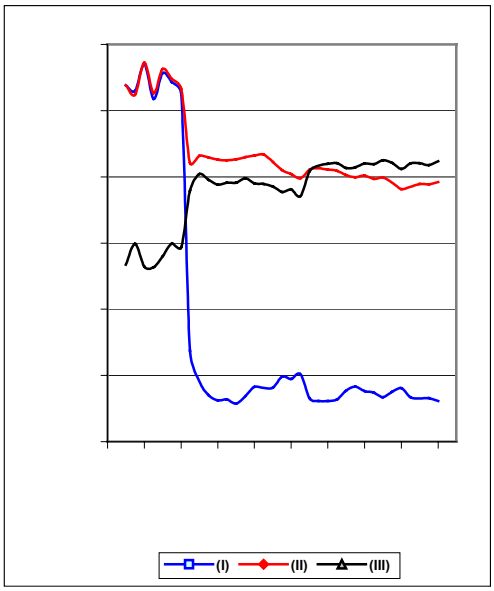
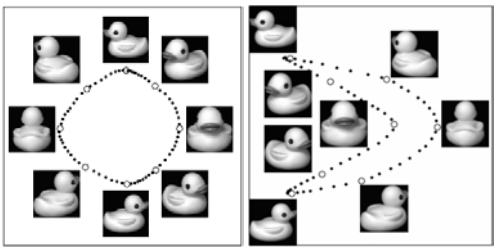


Figure 7. Dependences of Spearman's rho on k obtained after visualizing pictures of a rotating duckling by LLE:
(I) geodesic distances were evaluated in a space, $k_{geod} = 2$;
(II) geodesic distances were evaluated in a space, $k_{geod} = k$;
(III) - Euclidean distances were evaluated in a space



a) $k = 4$ b) $k = 9$

Figure 8. 2-dimensional embeddings of $m = 72$ pictures of a rotating duckling,
obtained by LLE using k nearest neighbours.
Larger circles mark representative samples of pictures

5. Conclusions

In this paper, we have explored the LLE algorithm for nonlinear dimensionality reduction. The main control parameter of LLE is the number of the nearest neighbours of each data point. This parameter greatly influences the results obtained. In this paper, we propose a new way for selecting the value of this parameter. In order to quantitatively estimate the topology preservation, we compute Spearman's rho.

The experiments have shown that the quantitative measure – Spearman's rho – is suitable to estimate the topology preservation after visualizing the data by the LLE algorithm. In order that Spearman's rho properly reflected the projections obtained, it is necessary to evaluate the geodesic but not Euclidean distances when calculating its value in an n-dimensional space by selecting rather a small number of neighbours in the geodesic distance algorithm.

Acknowledgment

The authors are very grateful to Dr. Olga Kayo and Dr. Oleg Okun from the Oulu University for their valuable remarks that allowed us to improve the quality of this paper.

The research is partially supported by the Lithuanian State Science and Studies Foundation project "Information technology tools of clinical decision support and citizens wellness for e.Health system (No. B-07019)".

References

[1] **C.C. Aggarwal, A. Hinneburg, D.A. Keim.** On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 1973, 2001.

[2] **M. Belkin, P. Niyogi.** Laplacian eigenmaps and spectral techniques for embedding and clustering. *In T.G. Dietterich, S. Becker and Z. Ghahramani (eds.), Advances in Neural Information Processing Systems* 14. *MIT Press*, 2002

[3] **C.M. Bishop, M. Svensén, C.K.I. Williams.** GTM: The generative topographic mapping. *Neural Computation*, 10(1): 215–234, 1998.

[4] **I. Borg, P. Groenen.** Modern multidimensional scaling. *Springer-Verlag, Berlin*, 1997.

[5] **T. Cox, M. Cox.** Multidimensional Scaling. *Chapman & Hall, London*, 1994.

[6] **T. Hastie and W. Stuetzle.** Principal curves. *Journal of the American Statistical Association*, 84: 502–516, 1989.

[7] **I.T. Jolliffe**. Principal Component Analysis. *Springer-Verlag, New York*, 1989.

[8] **T. Kohonen.** Self-organizing maps. *Springer Series in Information Sciences. Springer-Verlag, Berlin*, 1995.

[9] **O. Kouropteva, O. Okun, M. Pietikainen.** Selection of the optimal parameter value for the locally linear embedding algorithm. *Proc. of* 2002 *International Conference on Fuzzy Systems and Knowledge Discovery*, 2002, 359–363.

[10] **S. A. Nene, S. K. Nayar and H. Murase** Columbia Object Image Library (COIL-20). *Technical Report* CUCS-005-96, 1996.

[11] **S.T. Roweis and L.K. Saul.** Nonlinear dimensionality reduction by locally linear embedding. *Science,* 290, 2000, 2323–2326.

[12] **L.K. Saul, S.T. Roweis.** Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Machine Learning Research*, 4, *June* 2003, 119–155.

[13] **J.B. Tenenbaum, V. de Silva, J.C. Langford.** A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2000, 2319–2323.

Received September 2007.