

改进的线性局部切空间排列算法

李文华

(长江大学 计算机科学学院, 湖北 荆州 434023)

(wenhua999@qq.com)

摘要:线性局部切空间排列算法(LLTSA)是一种能很好地适用于识别问题的非线性降维方法,但LLTSA仅仅关注了数据的局部几何结构,而没有体现数据的整体信息。提出了一种基于主成分分析(PCA)改进的线性局部切空间排列算法(P-LLTSA),该算法在LLTSA的基础上,考虑了样本的全局结构,进而得到更好的降维效果。在经典的三维流形和在MNIST图像库手写体识别的实验中,识别率较PCA、局部保持投影算法(LPP),LLTSA有明显提高,证实了该算法在识别问题中的有效性。

关键词:主成分分析;局部切空间;流形学习;P-LLTSA算法;识别

中图分类号:TP39 **文献标志码:**A

Modified linear local tangent space alignment algorithm

LI Wen-hua

(College of Computer Science, Yangtze University, Jingzhou Hubei 434023, China)

Abstract: Linear Local Tangent Space Alignment (LLTSA) algorithm is a non-linear dimension reduction method which can be easily applied to recognition problems. It pays attention on the local geometric structure of data, but it neglects the global information of data. In this paper, an improved LLTSA algorithm based on principal component analysis (PCA) was proposed, and this method took the global structure of sample into consideration and contained a better reduction dimension result. In the classical experiment of 3D manifold and MNSIT image dataset script recognition, P-LLTSA has a higher recognition rate by contrast to PCA, LPP and LLTSA, which verifies the effectiveness of PLLTSA.

Key words: Principal Component Analysis (PCA); local tangent space; manifold learning; P-LLTSA algorithm; recognition

0 引言

随着计算机技术、多媒体技术、信息技术的飞速发展,海量数据及高维数据已经成为数据处理的一大难题。高维数据往往包含一些冗余维数,这些维数不但会降低数据处理的效率,还会增大数据处理的误差。这些高维数据同时也会表现出一定的线性或非线性的几何结构。如何降低这些数据的维数已经成为当今有挑战性的问题,目前已经有多种降维方法,如:主成分分析(Principal Component Analysis, PCA)^[1]、多维尺度变换^[2]等全局的线性降维算法。由于真实世界中的数据很多是非线性分布的,线性降维方法对这些数据的处理效果并不理想,所以近些年出现了很多基于流形学习的非线性降维算法。

流形学习的根本目的是希望找到嵌入在高维空间中潜在的低维流形结构,文献[3]在2000年首次提出了基于流形的局部降维算法——局部线性嵌入(Locally Linear Embedding, LLE)算法,首先寻找每个样本点的 K 个近邻点作为邻域,并希望每个样本点可以通过其近邻点近似线性表示,进而计算每个样本点邻域的局部重建权重矩阵,通过对该矩阵的特征分析得到低维嵌入。同年文献[4]在MDS的基础上通过计算最短路径来代替多维标度法(Multi-Dimensional Scaling, MDS)中的欧氏距离,提出了等距映射算法,而后文献[5]在LLE的基础上提出了局部线性坐标算法。在局部PCA的启

发下,文献[6]提出了局部切空间排列(Local Tangent Space Alignment, LTSA)算法。这些算法对非线性数据的降维都能得到比较理想的降维效果。然而由于其算法本身的局限性,这些算法都很难应用到识别领域。

为了解决这个问题,文献[7]提出了局部保持投影(Locality Preserving Projections, LPP)算法。通过对LLE的分析,文献[8]提出了一种基于LLE的特征分析算法,这些算法应用到识别问题中都能得到比较理想的效果。文献[9]提出了解决识别问题的新方法——线性的局部切空间排列(Linear-LTSA, LLTSA)算法,它给出了经典非线性降维算法LTSA的线性近似。但这些方法都仅考虑了整体数据集的局部信息,没有充分考虑到数据的整体信息。

本文通过对经典的非线性降维算法LTSA及其线性近似LLTSA的分析,提出了一种基于PCA及LLTSA新的降维算法,并将它应用于识别问题。本文的实验说明了该算法在识别问题中的有效性。

1 PCA和LTSA算法

1.1 PCA算法

PCA是一种经典的全局线性降维算法,它的主要思想是尽量保持高维空间嵌入到低维样本点的方差最大化,方差越大代表样本点的信息就越多,进而在降维的同时尽可能多地保留样本之间的信息,以下为PCA算法的描述。

考虑 n 个样本 $\{x_1, x_2, \dots, x_n\} \in \mathbf{R}^D$, 写成矩阵 $X = [x_1, x_2, \dots, x_n]$ 。我们要寻找高维坐标映射到低维子空间坐标 $T = [t_1, \dots, t_n]$ (其中 $T \in \mathbf{R}^{d \times n}$) 的一个线性变换 $U = [u_1, \dots, u_d] \in \mathbf{R}^{D \times d}$, 使得 $T = [u_1, \dots, u_d]^T X$, 并且 T 中的样本保持了方差最大。 n 个样本 $\{x_1, x_2, \dots, x_n\} \in \mathbf{R}^D$ 的协方差阵 $C_{D \times D} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})^T$ 其中 $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$ 。为了简写协方差阵, 我们把已经中心化的样本记为: $\hat{X} = [\hat{x}_1, \dots, \hat{x}_n]$, 其中 $\hat{x}_j = x_j - \bar{x}$, $\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j = \frac{1}{n} X e$, 其中 $e = [1, \dots, 1]^T$ 。于是 $\hat{X} = XJ$, 其中 $J = I - \frac{1}{n} e e^T$, 由此得样本的协方差矩阵为: $C_{D \times D} = \frac{1}{n-1} \hat{X} \hat{X}^T$ 。于是上述问题转化为如下的优化问题:

$$\begin{aligned} & \max \text{Tr}(U^T \hat{X} \hat{X}^T U) \\ \text{s. t. } & U U^T = I \end{aligned}$$

它等价于特征值问题: $\hat{X} \hat{X}^T u_i = \lambda_i u_i$, 其中 $i = 1, \dots, D$, 取前 d 个最大特征值对应的特征向量作为投影变换 U 。

1.2 LTSA 算法

局部追踪嵌入 (Locality Pursuit Embedding, LPE) 算法^[10]中证明局部的 PCA 特征空间就是局部样本均值的切空间, 因此可以对局部切空间进行运算。LTSA 是一种经典的非线性流形学习方法, 它主要有两个步骤, 即寻求最近邻点并实施局部 PCA 进行降维; 然后对所有的局部低维坐标整合到全局的低维坐标中, 进而到达非线性降维的目的。

考虑局部足够光滑的连续非线性映射 (至少是一阶光滑) $f = (f_1, \dots, f_D)^T$, 任意低维向量 $t = (t_1, t_2, \dots, t_d)^T$, \bar{t} 为局部样本点的均值, 则在点 \bar{t} 处, 有泰勒展开式: $f(t) = f(\bar{t}) + J_{\bar{t}}(t - \bar{t}) + O(\|t - \bar{t}\|_2^2)$, $J_{\bar{t}}$ 为 f 的雅克比矩阵, 也就是 d 维切空间的一组标准正交基, 它可由局部 PCA 特征矩阵近似代替, 当 t 在 \bar{t} 足够小的邻域中时, 近似地, 可写成 $f(t) = f(\bar{t}) + J_{\bar{t}}(t - \bar{t})$ 。

对高维空间的数据点 $x_i = f(t_i) + \varepsilon_i$ 。在点 x_i 处的 k 最近邻邻域中实施 PCA, $X_i = [x_{i1}, x_{i2}, \dots, x_{ik}]$ 为包含本身的 k 个最近邻点所构成的矩阵。由上述讨论过程, 有:

$$x_{ij} - \bar{x}_i = f(t_{ij}) - f(\bar{t}_i) = J_{\bar{t}_i}(t_{ij} - \bar{t}_i) = Q_i \theta(t_{ij}) \quad (1)$$

其中 Q_i 为在该邻域应用 PCA 所得的样本切空间的基底, $\theta(t)$ 为其对应切空间的局部坐标。进而得到 $\theta_{ij} = Q_i^T(x_{ij} - \bar{x}_i)$ 。因此有: $x_{ij} = \bar{x}_i + Q_i \theta_{ij} + \varepsilon_{ij}$ 。其中 $\varepsilon_{ij} = (I - Q_i Q_i^T)(x_{ij} - \bar{x}_i)$ 为重构误差。

下面考虑怎样由得到的有局部几何特征的 $\{\theta_{ij}\}$, 构造低维特征空间中全局坐标 $t_i, i = 1, 2, \dots, n$ 。定义 $\Theta_i = [\theta_{i1}, \dots, \theta_{ik}]$, $T_i = [t_{i1}, \dots, t_{ik}]$, 极小化重建误差:

$$\min_{t_i, L_i} \sum_{i=1}^n \| \varepsilon_i \|_2^2 = \min_{t_i, L_i} \sum_{i=1}^n \| T_i - \bar{t}_i e^T - L_i \Theta_i \|_2^2 \quad (2)$$

其中 \bar{t}_i 为 t_{i1}, \dots, t_{ik} 的均值, L_i 为待定的局部仿射变换矩阵。对 T_i 固定, 使得重建误差极小的 $L_i = (T_i - \bar{t}_i e^T) \Theta_i^+ = T_i \Theta_i^+$ 。其中 Θ_i^+ 为 Θ_i 的 Moore-Penrose 广义逆。因此, 该最优问题等价于:

$$\min_f \| T S W \|_F^2 \quad (3)$$

其中 $S = [S_1, S_2, \dots, S_n]$, S_i 为满足 $T S_i = T_i$ 的 0-1 选择矩阵, $W = \text{diag}(W_1, \dots, W_n)$, 其中 $W_i = (I - \frac{1}{k} e e^T)(I - \Theta_i \Theta_i^+)$ 。

为了唯一确定 T , 施加规范性约束 $T T^T = I_d$ 。显然 e_d 是矩阵 $B = S W W^T S^T$ 的相应零特征值对应的特征向量。(已经证明) 因此最优的 T 由 B 的第 2 到第 $d+1$ 个最小特征值的特征向量 u_2, \dots, u_{d+1} 给出 $T = [u_2, \dots, u_{d+1}]^T$ 。

由于 LTSA 是一种隐式的非线性映射, 它对非线性流形的降维效果很好, 但对于稀疏样本分布或者高曲率样本分布效果都不是很理想, 且很难应用到识别领域, 所以下面介绍线性的 LTSA (LLTSA), 它的提出解决了 LTSA 的识别问题。

2 基于 LTSA 的线性识别的 P-LLTSA 算法

LLTSA 受到 LTSA 的启发, 它依然首先执行局部 PCA, 然后将局部低维坐标整合到全局的低维子空间。在式 (3) 的基础上借用 PCA 的思想将线性映射 $T = U^T \hat{X}$ 代入到式 (3) 及规范性约束中, 得到 LLTSA 的优化模型:

$$\begin{aligned} & \min U^T X J_n B J_n X^T U \\ \text{s. t. } & U^T X J_n X^T U = I_d \end{aligned} \quad (4)$$

最终的优化问题转化成为一个广义特征值问题: $X J_n B J_n X^T u_i = \lambda_i X J_n X^T u_i, i = 1, \dots, D$ 。将特征值排序 $\lambda_1 < \lambda_2 < \dots < \lambda_D$, 取前 d 个最小特征值对应的特征向量作为最后的变换 U , 进而得到 $T = U^T \hat{X}$ 。

LLTSA 充分考虑了高维样本点的局部信息和拓扑结构, 但对整体信息缺少描述, 因此, 本文提出一种同时兼顾到局部和整体样本信息的新算法——P-LLTSA 算法。该算法的主要思想是保证 LLTSA 局部重构误差之和最小的同时, 使整体的样本点的方差最大化, 基于这个思想, 将其描述为以下优化问题:

$$\begin{aligned} & \min \frac{U^T X J_n B J_n X^T U}{U^T X J_n X^T U} \\ \text{s. t. } & U^T X J_n X^T U = I_d \end{aligned} \quad (5)$$

式 (5) 等价于一个广义特征值问题: $X J_n B J_n X^T u_i - X J_n X^T u_i = \lambda_i X J_n X^T u_i$, 由于 J_n 是一个幂等矩阵, 则上面的特征问题可写成 $X J_n (B - I) J_n X^T u_i = \lambda_i X J_n X^T u_i$ 。

由于该算法在计算广义特征问题时出现了 $B - I$, 这就增加了该矩阵奇异的可能性, 为了解决这个问题, 将原有的优化模型修正为:

$$\begin{aligned} & \min \frac{(1 + \alpha) U^T X J_n B J_n X^T U}{\alpha U^T X J_n X^T U} \\ \text{s. t. } & U^T X J_n X^T U = I_d \end{aligned} \quad (6)$$

其中 $0 < \alpha < 1$, 它可以减小特征值问题出现奇异现象的可能。

在识别问题中很多情况下样本维数大于样本数, 导致出现小样本问题, 因此采用如下的识别算法。

算法 1 P-LLTSA 算法。

步骤 1 投影。对于已经向量化的 n 个样本的数据集 X , 选取适当的维数 d , 使得 $d < n$, 在样本空间利用 PCA 进行投影, 计算投影矩阵 $U_{\text{PCA}}, X \leftarrow U_{\text{PCA}} X$ 。

步骤 2 寻找邻域。对于每个样本 $x_i, i = 1, 2, \dots, n$, 寻找 x_i 的 k 最近邻点, 作为其邻域。

步骤 3 提取局部信息。计算矩阵 $X_i H_k$ 的右奇异向量组 V_i , 令 $W_i = H_k (I - V_i V_i^T)$ 。

步骤 4 初始化特征矩阵 $B = 0$, 及邻域指标集 $I_i = \{i_1, \dots, i_k\}$ 为 x_i 的 k 最近邻指标, 对矩阵 B 进行如下更新: $B(I_i, I_i) \leftarrow B(I_i, I_i) + W_i W_i^T, i = 1, 2, \dots, n, B = B - I$ 。

步骤 5 按照式 (5) 计算嵌入; 计算广义特征值问题: $X J_n (B - I) J_n X^T u_i = \lambda_i X J_n X^T u_i$ 。

将其特征值排序 $\lambda_1 < \lambda_2 < \dots < \lambda_d$, 取前 d' 个最小特征值对应的特征向量作为最后的变换 U , 进而得到 $T = U^T U_{PCA}^T \hat{X}$ 。

3 实验结果

3.1 三维流形

在三维空间选取 1 000 个样本, 采用 swiss-roll、twin-peaks、3D-clusters 这三种三维流形, 分别使用 PCA、LLTSA、P-LLTSA 三种算法得到的二维嵌入如图 1~3 所示。

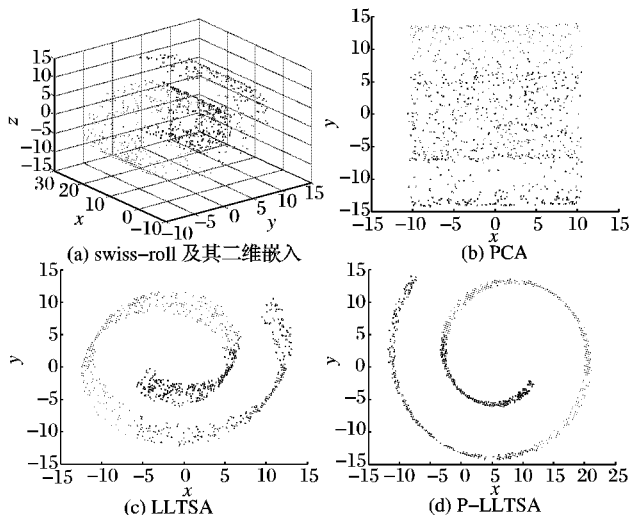


图1 swiss-roll 三维流形及其二维嵌入

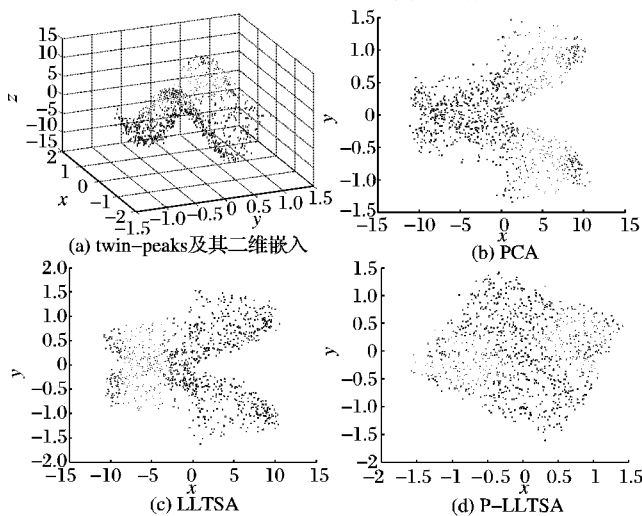


图2 twin-peaks 三维流形及其二维嵌入

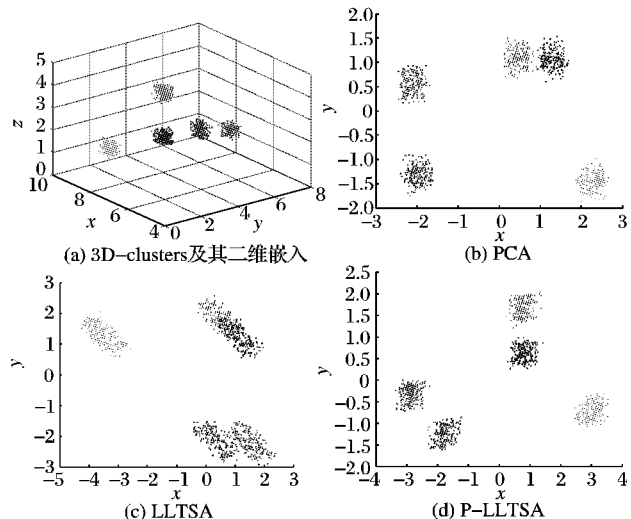


图3 3D-clusters 三维流形及其二维嵌入

从三维流形的二维嵌入效果可以看出, P-LLTSA 的二维嵌入要优于 PCA 及 LLTSA, 尤其 3D-clusters 数据集中尤为明显, PCA 对蓝色和褐色的二维嵌入有少量重叠, LLTSA 有一部分重叠, 只有 P-LLTSA 成功将五种 3D-clusters 成功区分开。

3.2 手写体识别

采用 MNIST 手写体数据库, 共有 0~9 手写体数字, 每类约 6000 个样本作为训练集, 约 1000 个作为测试集, 本文随机选取三个数字进行识别实验 (如图 4), 在训练集中选取 20、30、40、50 个训练样本, 采用最小距离分类器, 用 20 个测试样本进行测试, 得到手写体数字的识别率随维数变化曲线如图 5 所示。



图4 随机选取三类(每类 10 个样本)手写体训练样本

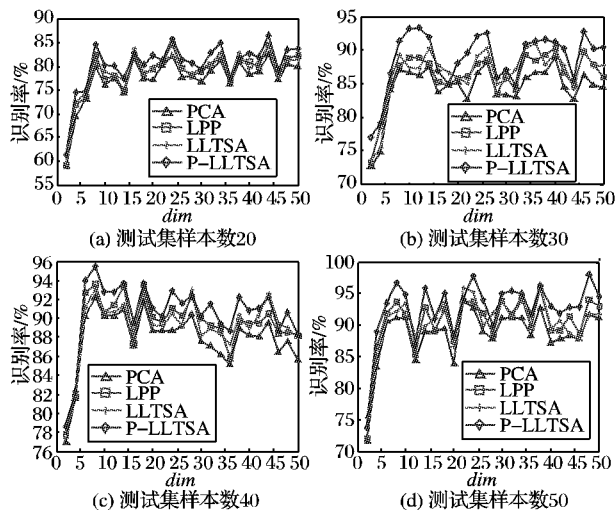


图5 手写体数字的识别率随维数 dim 变化曲线

从图 5 中可以看到 2~50 维 PCA、LPP、LLTSA 和 P-LLTSA 对应的识别率, 为了减小误差, 选取第二个最高识别率作为某种方法应用在手写体识别的最高识别率, 记为 R_{max} 。表 1 为各种算法 R_{max} 及维数的对比。从表 1 中可以很容易看出 P-LLTSA 的 R_{max} 高于其他三种算法, 而且图 5 中也显示 P-LLTSA 的识别率也普遍高于另外三种算法。

表1 手写体不同测试集最高识别率及对应维数

测试集	20samples		30samples		40samples		50samples	
	dim	$R_{max}/\%$	dim	$R_{max}/\%$	dim	$R_{max}/\%$	dim	$R_{max}/\%$
PCA	38	80.17	26	88.15	18	91.78	22	92.78
LPP	44	84.53	10	88.85	8	93.63	22	94.20
LLTSA	16	83.78	14	90.36	18	92.66	22	95.93
P-LLTSA	24	85.80	10	93.21	6	93.96	24	97.78

利用最大似然估计 (Maximum Likelihood Estimation, MLE)^[11] 本征维数估计方法在 MNIST 库中进行多次实验, 由于此类本征维数估计方法都是基于样本点密集提出的, 所以每次随机选取 2000 幅 0~9 中的两类图像, 得到手写体的本征维数集中在 6~11 维, 这与表 1 中 P-LLTSA 中 30 及 40 测试样本的最高识别率相吻合。由此可见该算法的有效性。

4 结语

本文提出了一种改进的线性局部切空间排列算法, 该算

(下转第 253 页)

中数据分布均呈圆形(如图8)。

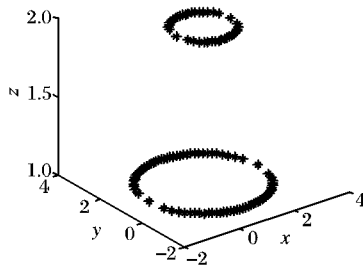


图8 人为构造的两类三维空间数据集

用 ILDA 对上述数据集进行降维后得到图9。不难发现两类数据被明显分开,可以说 ILDA 能很好地发现数据的内在特征。

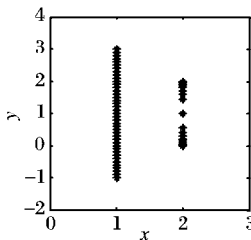


图9 三维空间数据集降维效果示意图

5 结语

本文在 Fisher 准则基础上,提出一种改进的 LDA 算法 ILDA。该方法引入了类间离散度标量和类内离散度标量,使得最佳鉴别方向的确定不受类内离散度矩阵奇异的制约,也不受限于类间离散度矩阵的秩。标准的 ORL 人脸数据库和人工数据集进行的实验表明:尽管 ILDA 的识别率略低于 LDA,但 ILDA 突破了 LDA 秩限制问题,可以提取远大于 $c-1$ 个鉴别特征。与此同时,ILDA 使得计算量大幅降低,运算效率有所提高。

参考文献:

- [1] BELHUMEUR P N, HESPANHA H P, KRIEGMAN D J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.
- [2] YU HUA, YANG JIE. A direct LDA algorithm for high-dimensional data with application to face recognition [J]. Pattern Recognition, 2001, 34(10): 2067-2070.
- [3] 杨健. Fisher 线性鉴别分析的理论研究及其应用 [J]. 自动化学报, 2003, 29(4): 481-493.
- [4] YANG JIAN, YANG JINGYU. Why can LDA be performed in PCA transformed space [J]. Pattern Recognition, 2003, 36(2): 563-566.
- [5] TIAN Q, FAJMAN Y, LEE S H. Comparison of statistical pattern recognition algorithms for hybrid processing [J]. Journal of the Optical Society of America, 1988, 5(10): 1655-1669.
- [6] LIU KE, CHENG YONGQING, YANG JUYANG, et al. An efficient algorithm for Foley-Sammon optimal set of discriminant vectors by algebraic method [J]. International Journal of Pattern Recognition and Artificial Intelligence, 1992, 6(5): 817-829.
- [7] ZHONG JIN, YANG JINYU. Face recognition based on uncorrelated discriminant transformation [J]. Pattern Recognition, 2001, 34(7): 1405-1416.
- [8] 金忠, 杨靖宇. 一种具有统计不相关性的最优鉴别矢量集 [J]. 计算机学报, 1999, 22(10): 1105-1108.
- [9] SWEET D, WENG J. Using discriminant eigenfeatures for image retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(8): 831-836.
- [10] TURK M. A pentland. eigenfaces for recognition [J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.
- [11] TOSHIHKO O, SHINGO T. An optimal orthonormal system for discriminant analysis [J]. Pattern Recognition, 1985, 18(2): 139-144.
- [12] HONG ZIQUAN, YANG JINGYU. Optimal discriminant plane for a small number of samoles and design method of classifier on the plane [J]. Pattern Recognition, 1991, 24(4): 317-324.
- [13] 李道红. 线性判别分析新方法研究及其应用 [D]. 南京: 南京航空航天大学, 2005.
- [14] FRIEDMAN J. Regularized discriminant analysis [J]. Journal of the American Statistical Association, 1989, 84(405): 165-175.
- [15] 杨健, 杨静宇, 叶晖. Fisher 线性鉴别分析的理论研究及其应用 [J]. 自动化学报, 2003, 29(4): 481-493.
- [16] AT&T Laboratories Cambridge. The ORL database of faces [EB/OL]. [2010-02-10]. <http://www.cam-orl.co.uk/facedatabase.html>
- [17] 李刚, 高政. 人脸识别理论研究进展 [J]. 计算机与现代化, 2003(5): 1-6.

(上接第249页)

法在 LLTS 算法的基础上,利用 PCA 算法丰富了样本分布的整体信息,补充了 LLTS 缺少整体信息的缺点,并在三维流形和手写体识别实验中证实了算法的优越性。此外,该方法的修正模型减小了 LLTS 在样本相对较少时特征分解出现奇异现象的可能,也没有额外增加算法的时间复杂度。但是该算法在样本较少时还不够稳定,需要在此基础上设计一种更加稳定的算法。

参考文献:

- [1] JOLLIFFE I T. Principal component analysis [M]. New York: Springer-Verlag, 1986.
- [2] COX T, COX M. Multidimensional scaling [M]. London: Chapman and Hall, 1994.
- [3] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.
- [4] TENENBAUM J B, de SILVA V, LANGFORD J. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(5500): 2319-2323.
- [5] TEH Y W, ROWEIS S T. Automatic alignment of hidden representations [C]// Advances in Neural Information Processing Systems. Cambridge: MIT, 2002: 841-848.
- [6] ZHANG ZHEN-YU, ZHA HONG-YUAN. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment [J]. SIAM Journal of Scientific Computing, 2004, 26(1): 313-338.
- [7] HE XIAO-FEI, NIYOGI P. Locality preserving projections [EB/OL]. [2010-02-15]. http://books.nips.cc/papers/files/nips16/NIPS2003_AA20.pdf.
- [8] FU YUN, HUANG T S. Locally linear embedded eigenspace analysis [D]. Champaign: University of Illinois at Urbana-Champaign, 2005.
- [9] ZHANG TIANHAO, YANG JIE, ZHAO DELI, et al. Linear local tangent space alignment and application to face recognition [J]. Neurocomputing, 2007, 70(7/9): 1547-1553.
- [10] MIN WANLI, LU KE, HE XIAOFEI. Locality pursuit embedding [J]. Pattern Recognition, 2004, 37(4): 781-788.
- [11] LEVINA E, BICKEL P J. Maximum likelihood estimation of intrinsic dimension [C]// Advances in Neural Information Processing Systems. Cambridge: MIT, 2005: 51.