基于软K段主曲线算法的字符特征提取研究及实现*)

焦 娜[1,2] 迟呈英[2] 苗夺谦[1] 杨 红[3]
(鞍山科技大学计算机科学与工程学院 辽宁 114002)[1] (同济大学计算机科学与工程系 上海 200092)[2]
(国家高性能计算机工程中心同济分中心 上海 200092)[2] (鞍山市科学技术情报研究所 辽宁 114001)[3]

摘 要 要提高脱机手写字符识别的识别率,关键是特征的提取。主曲线是主成分分析的非线性推广,是通过数据分布"中间"并满足"自相合"的光滑曲线。通过对现有主曲线算法分析可知:软K段主曲线算法对提取出分布在弯曲度很大或相交曲线周围的数据的主曲线效果较好。因此本文尝试用该主曲线算法来提取脱机手写字符的结构特征。实验结果表明,利用该主曲线算法来提取脱机手写字符的结构特征不但是可行的,而且取得较好的实验效果。它为脱机手写字符特征提取的研究提供了一条新途径。

关键词 软K段主曲线算法,结构特征,特征选取

Research and Implementation of Structural Features of Characters Based on a Soft K-Segments Algorithm for Principal Curves

JIAO Na[1,2] CHI Cheng-Ying[1] MIAO Duo-Qian[2] YANG Hong[3]
(School of Computer Science and Technology, Anshan University Science and Technology, Liaoning 114002)[1]
(Department of Computer Science and Technology, Tongji University, Shanghai 200092)[2]
(Tongji Branch, National Engineering & Technology Center of High Performance Computer, Shanghai 200092)[2]
(Anshan Scientific Technological Information Institute, Liaoning 114001)[3]

Abstract Extraction of features is critical to improve the recognition rate of off-line handwritten characters. Principal curves are nonlinear generalizations of principal components analysis. They are smooth self-consistent curves that pass through the "middle" of the distribution. By analysis of existed principal curves, we learn that a soft k-segments algorithm for principal curves exhibits good performance in such situations in which the data sets are concentrated around a highly curved or self-intersecting curves. Therefore, we attempt to use the algorithm to extract structural features of off-line handwritten characters. Experiment results show that the algorithm is not only feasible for extraction of structural features of characters, but also exhibits good performance. The proposed method can provide a new approach to the research for extraction of structural features of characters.

Keywords Soft K-segments algorithm for principal curves, Structural features, Features extraction

1 引言

主曲线概念[1]是 Hastie 和 Stuetzle 于 1984 年提出的。主曲线是通过数据分布"中间"并满足"自相合"的光滑曲线,其理论基础是寻找嵌入高维空间的非欧氏低维流形,也是线性主成分的非线性推广[7]。由于主曲线的这些性质和优点,自上世纪 90 年代以来在国外取得了较快的发展。1992 年 Banfield 和 Raftery 提出了 BR 主曲线[2];1999 年,Kegl 提出了 PL 主曲线[3];2000 年,Verbeek 给出了 K 段主曲线算法[4];2001 年,Delicado 提出了 D 主曲线[6]。虽然在主曲线的原理中使用了较复杂的数学,但由于其广泛的应用前景,在 90 年代后期已引起国外计算机科学家的关注,现在他们已报道了许多主曲线在计算机方面的应用,如线性对撞机中对电子束运行轨迹的控制、图像处理中辨识冰原轮廓、脱机手写体的主曲线模板化和数据可听化等。

由于 HS 主曲线算法、PL 主曲线算法、BR 主曲线算法和 T 主曲线算法共同存在的问题是:它们都是由与固定拓扑结构相关的局部模型的组合组成的,因此当数据分布在弯曲度很大或相交曲线周围时,这些算法的性能就差,所得结果不能正确反映数据的拓扑结构。这是由在"局部模型"中的固定拓扑结构和差的初始化造成的。因为我们预先不知道需要多少"局部模型",所以在设计算法时,只能估计"局部模型"的数量,另一方面 HS 算法明确规定主曲线是不相交的。同样,多边形算法和 T 算法也由于其差的初始化,使得它们通常不能正确提取出分布在弯曲度很大或相交曲线周围的数据的主曲线(如图 1 所示)。脱机手写字符图像具有弯曲度大和相交等特点,所以 HS 主曲线算法、多边形算法和 T 算法都不适用来提取脱机手写字符的骨架结构。由于软 K 段主曲线算法[5]对提取出分布在弯曲度很大或相交曲线周围的数据的主曲线效果较好,根据软 K 段主曲线算法的这个特点,我们尝试选用它来提取脱机手写字符的骨架结构。实验结果表明,利用该主曲线来提取脱机手写字符的结构特征不但是可行的,而且取得了较好的实验效果。它为脱机手写字符特征提取的研究提供了一条新途径。

*)基金项目:国家自然科学基金项目(60175016,60475019)。焦 娜 硕士研究生。

图 1　多边形算法和 HS 算法提取数据的主曲线

2　主曲线定义及软 K 段主曲线算法

这一部分我们给出主曲线的定义以及软 K 段主曲线算法。

2.1　主曲线定义

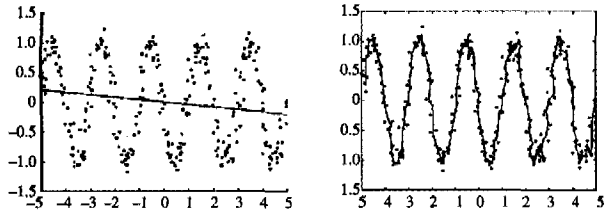
定义 1　假设随机向量 $Y=(Y_1,Y_2,\cdots,Y_p)$ 的概率密度为 $g_y(y)$，则通过 Y 数据分布中间的一条曲线 $f(s)$ 如果满足：

$$f(s)=E(Y|s_f(y)=s)$$

则称 $f(s)$ 是 Y 的一条主曲线。其中 $s_f(y)$ 是数据点 y 投影到曲线 $f(s)$ 上 s 点的值，即

$$s_f(y)=\sup\{s:\|y-f(s)\|=\inf_{\tau}\|y-f(\tau)\|\}$$

由主曲线定义可知：主曲线上每个点是所有投影至该点的数据点的条件均值，它满足自相合性。主曲线的理论基础是寻找嵌入高维空间的非欧氏低维流形，也是线性主成分的非线性推广，能真实反映数据的形态。图 2 是一个简单的例子，从该图中可发现主曲线与第一主成分相比具有两个明显的优点：一方面对数据的信息保持性好，另一方面它与数据间的距离均方差小，较好地勾画出了原始信息的轮廓。



(a)第一主成分线　　(b)主曲线

图 2　数据第一主成分与主曲线的对照图

2.2　软 K 段主曲线算法

算法主要由以下几步组成：

① 初始化步。初始化分为两步：读入数据点集 $X=(x_1,x_2,\cdots,x_n)$ 和计算其第一主成分线，然后从中间取 3σ 作为初始线段的长度，σ^2 是第一主成分线的方差，令初始线段为 s_1，s_1 的 Voronoi 区域为 $V_1=\{x_1\cdots x_n\}$，$k=1$。

② 插入一条新线段。判断 k 是否小于 $k\_\max$。如果 k 不小于 $k\_\max$，则程序结束，否则计算点 x_q，该点满足：

$$x_q=\inf\{x_t:|\sum_{i=1}^{n}g(x_i,x_t)=\max(\sum_{i=1}^{n}g(x_i,x_j))\}$$

其中

$$g(x_i,x_j)=\begin{cases}dist(x_i)-d(x_i,x_j) & dist(x_i)-d(x_i,x_j)>0\\ 0 & dist(x_i)-d(x_i,x_j)\leqslant 0\end{cases}$$

$1\leqslant i,j\leqslant n$，在 $g(x_i,x_j)$ 中 $dist(x_i)=\min_{j=1,2,\cdots,k}d(x_i,s_j)$，$d(x_i,x_j)=\|x_i-x_j\|^2$，则求出点 x_q 的 Voronoi 区域。Voronoi 区域为：

$V_q=\{x\in X|\|x-x_q\|\leqslant\min d(x,s_j),j=1,2,\cdots,k\}$ 然后求 V_q 的第一主成分线，从该线中间取 3σ 长作为新插入线段，σ^2 是第一主成分线的方差。$k=k+1$，令新插入线段为 s_k，s_k 的 Voronoi 区域为 $V_k=\Phi$。

③ 调整步。调整新线段与其它线段。具体算法如下：

1)设每条线段旧的 Voronoi 区域$(V_1,V_2,\cdots,V_k)$。

2)求出每条线段新的 Voronoi 区域。$\forall s_i,i=1,\cdots,k$ 求

$$V'_i=\{x_j\in X|\|x_j-s_i\|=\min_{i=1,2\cdots,k}\|x_j-s_t\|\}$$

3)比较$(V'_1,V'_2,\cdots,V'_k)$与$(V_1,V_2,\cdots,V_k)$是否相同。如果不同，求出所有 $V'_j(j=1,2,\cdots,k)$的第一主成分线，并且把$(V'_1,V'_2,\cdots,V'_k)$赋给$(V_1,V_2,\cdots,V_k)$，继续第 2 步；如果相同，调整步结束。

④ 构造优化步。将 k 条线段构造成一条哈密尔顿路径，并进行优化。算法如下：

1)令 $p=k$(p 为子哈密尔顿路径的个数)，则 p 个子哈密尔顿路径有 $2p$ 个端点、2^p 个边。

2)如果 $p<2$，停止。否则，求 2^p 个边的代价值 $c(e_i)$，其中 $c(e_i)=l(e_i)+\lambda a(e_i)$；$e_i=(v_l,v_m)$，$(v_l,v_m$ 分别是两个不同子哈密尔顿路径的端点)；$0\leqslant\lambda\in R$ 是用户定义的参数；$l(e_i)$为边 e_i 的长度，$a(e_i)$是角度惩罚，$a(e_i)=\alpha+\beta$(如图 3 所示)，连接使 $c(e_i)$最小的边的端点，$p=p-1$，返回第 2 步。

3)用 $2-opt$ 的 TSP(城市推销员问题)优化方案来优化所形成的 HP(哈密尔顿路径)。

4)计算目标函数 $OF=n\log l+\sum_{i=1}^{k}\sum_{x\in V_I}d(s_i,x)^2/(2\sigma^2)$，$l$ 为构造后的哈密尔顿路径长度。如果 OF 最小，则程序结束，否则返回第 2 步。

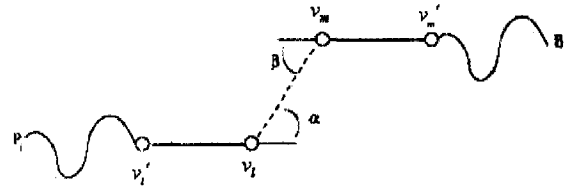


图 3　子哈密尔顿路径角度惩罚图

软 K 段主曲线算法提取脱机手写数字的程序流程见图 4。

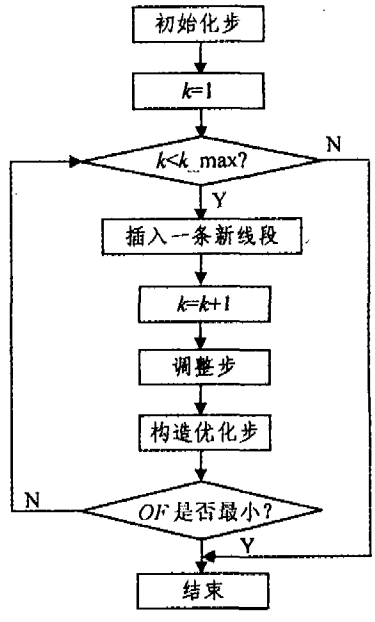


图 4　软 K 段主曲线算法流程图

3 基于软K段主曲线算法的脱机手写字符特征提取

本节我们把软K段主曲线算法应用在脱机手写字符特征提取上。我们利用六种脱机手写字符作为实例来进行实验。利用2.2节中软K段主曲线算法得到的主曲线，结果如图5所示。

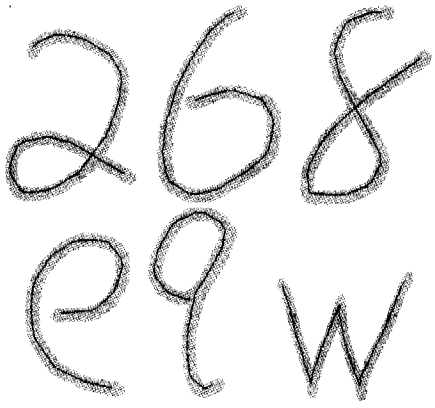


图5 软K段主曲线算法提取脱机手写字符骨架图

从图5可以看出，软K段主曲线算法能够很好地应用在弯曲度很大或相交脱机手写字符上。由实验我们可知，每一个脱机手写字符由几条曲线组成：一条曲线是由一系列点$(v_{i_1},\cdots,v_{i_l})$组成，并且每对相邻的点$(v_{i_j},v_{i_{j+1}})$，$j=1,\cdots,l-1$有一条边相连。例如第一个字母由3条曲线组成。

由于脱机手写字符的书写因人而异，因时而变，形态变化十分巨大。为了更有利于有效鉴别特征的选取，我们在用软K段主曲线算法提取出脱机手写字符的主曲线特征后，在进行特征选取前先对近似回路和回路外的短分支和小圈进行预处理，再进行有效模式特征选取。通过对脱机手写字符主曲线特征详细分析后发现：笔画数和回路数是区分脱机手写字符的重要特征，且易于检测。这里我们把脱机手写字符中光滑曲线段的个数称之为笔画数(一个回路表示一画)。我们知道，对同一脱机手写字符，写法不同，笔画数也可能不同，例如L和l、9和g、m和m、h和h、2和2等。考虑到这个原因，首先我们把笔画数作为整体特征来进行分类，然后对笔画数相同的每一类，用回路数这一整体结构特征来对脱机手写字符进行进一步粗分类。例如笔画数为2时，可分三类{abdegpq69}、{8}、{ijmnvwz23}；最后，我们将回路数、水平线数、竖直线数、是否为直线、水平线端点相对于交点的位置、凹/凸点相对于交点的位置、主要曲线相对于回路的位置、凸点相对于凹点的位置、竖直线端点相对于交点的位置、用回路中心相对于交点位置、最大分叉数、最大分叉数与笔画数的关系、是否存在点、竖直线数、右凸点数、上凸点数、左凹点数、下凹点数、笔画端点相对于回路的位置作为其细节特征将脱机手写字符进一步区分。例如，我们通过分析发现通过回路中心相对于交点位置和笔画端点相对于回路的位置就可以把{abdegpq69}区分开；W不管如何写，预处理后其下凹点数必为2；通过凸点和凹点的相对位置可区分S和Z等规则。3不管如何写，其右凸点数必为2。

结论 本文提出了用软K段主曲线算法来找到数据集的主曲线。用此方法对弯曲度很大或相交的曲线效果较好。并把此方法应用在脱机手写字符特征提取上。从实验结果看，该算法对大部分弯曲度很大或相交脱机手写字符都取得了较好的效果，但还有一些不尽人意的地方，需要我们在下一步的工作中进行改进。通过对所做实验进行分析，我们认为把软K段主曲线算法应用在脱机手写字符特征提取上是可行的。我们计划在今后的研究中完善算法，以便能够取得更好的效果，同时还要将此算法与其他算法进行比较，分析它与其他算法的效率和性能上的优劣。

参考文献

1 Hastie T. Principal curves and surfaces:[Technical Report]. Laboratory for Computational Statistics, Stanford University, Department of Statistics, 1984

2 Banfield J D, Raftery A E. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Journal of the American Statistical Association, 1992, 87(417):7~16

3 Kegl B, Krzyzak A, et al. A polygonal line algorithm for constructing principal curves. In: Proceedings of Neural Information Processing System, 1999. 501~507

4 Verbeek JJ, Vlassis N, Krose B. A k-segments algorithm for finding principal curves. Pattern Recognition Letters, 2002, 23(8): 1009~1017

5 Verbeek JJ, Vlassis N, Krose B. A soft k-segments algorithm for principal curves:[Technical Report]. Computer Science of Institute, University of Amsterdam, ICANN 2001. 450~456

6 Delicado P. Another look at principal curves and surfaces. Journal of Multivariate Analysis, 2001, 77(1):84 ~116

7 张军平，王珏. 主曲线综述. 计算机学报，2003，26(2)：129~146

8 唐庆适，苗夺谦，张红云. 基于主曲线进行指纹细节特征提取的方法[J]. 计算机科学，2005(1)