计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/ CN 11-1777/ TP
42(8): 1344～1349, 2005

基于主曲线的脱机手写数字结构特征分析及选取

张红云　苗夺谦　张东星
(同济大学计算机科学与技术系　上海　200092)
(zhanghongyun583 @sina.com)

Analysis and Extraction of Structural Features of Off-Line Handwritten Digits Based on Principal Curves

Zhang Hongyun, Miao Duoqian, and Zhang Dongxing
(*Department of Computer Science and Technology, Tongji University, Shanghai* 200092)

Abstract　Extraction and choice of features are critical to improving the recognition rate of off-line handwritten digits. Principal curves are nonlinear generalizations of principal components analysis. They are smooth self-consistent curves that pass through the "middle" of the distribution. They preferably reflect the structural features of the data. During digit feature selection, firstly principal curves are used to extract the structural features of training data; Secondly the classification features used for digits coarse classification and precise classification are chosen by analyzing the structural features of principal curves in detail; Finally coarse classification and precise classification are separately carried out in handwritten digits recognition. The Concordia University CENPARMI handwritten digit database is used in the experiment. The result of the experiment shows that these features have good discriminating power of similar digits. The proposed method can effectively improve the recognition rate of off-line handwritten digits and provide a new approach to the research for off-line handwritten digits recognition.

Key words　principal curves; structural features; features extraction

摘　要　要提高脱机手写数字识别的识别率,关键是特征的提取与选择. 主曲线是主成分分析的非线性推广,它是通过数据分布"中间"并满足"自相合"的光滑曲线. 它较好地反映了数据分布的结构特征. 在数字特征选取中,首先将主曲线用于训练数据的特征提取;其次在详细分析数字主曲线的结构特点的基础上,选择出用于数字识别的粗分类、细分类特征;最后在对手写数字进行识别时,先进行粗分类再进行细分类. 所提方法在 Concordia 大学的 CENPARMI 手写体数字数据库上的实验结果表明:利用这些特征能有效区分相似字符,提高了手写数字的识别率,为脱机手写数字识别的研究提供了一条新途径.

关键词　主曲线; 结构特征; 特征选取

中图法分类号　TP391.41

1　引　　言

众所周知,影响计算机模式识别系统性能好坏的关键因素之一是模式特征的提取与选择. 对手写数字识别,目前有效的特征描述量有轮廓、骨架、KL 变换、Fourier 变换、方向线素、矩、网格、投影、小波变换等. 针对不同的应用,特征的选择也不尽相同,但目的都是降低识别的误识率与拒识率. 与上面描述量相比,主曲线能很好地描述模式特征. 它具有信息保持性好、维数低、平移和伸缩不变性等优点;并且较其他结构特征相比,从图论的角度看,它更有利于进行有效特征的选择.

主曲线概念[1]是 Hastie 于 1984 年提出的. 主

收稿日期:2003-11-04;**修回日期**:2005-04-07
基金项目:国家自然科学基金项目(60175016,60475019);国家"九七三"重点基础研究发展规划基金项目(2003CB316902);上海市科委重大科技攻关基金项目(03DZ15029)

曲线是通过数据分布“中间”并满足“自相合”的光滑曲线,其目的是根据给定的数据集合求出一条曲线,使得这条曲线对给定的数据集合是某种意义下的对偶. 形象地说,希望能寻找通过数据分布“中间”的曲线,使它能真实地反映数据的形态,即曲线是数据集合的“骨架”,数据集合是这个曲线的“云”. 由此可见,主曲线对数据的信息保持性好. 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[2]. 由于主曲线的这些性质和优点,自 20 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[3],1999 年 Kegl 等人提出了 PL 主曲线[4],2000 年 Verbeek 等人给出了 K 段主曲线算法[5],2001 年 Delicado 等人提出了 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 20 世纪 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、手写体的主曲线模板化和数据可听化等.

本文首先采用推广的多边形(PL)主曲线算法来提取手写数字的骨架结构[7];然后在详细分析数字主曲线结构特征的基础上,选择数字图像中的回路数作为其整体特征,将笔画数、是否为直线、凸曲线数、凹曲线数、凸凹点及交点的相对位置、笔画端点相对于回路的位置作为其细节特征;最后将两者分别用于对手写数字进行粗分类和细分类. 所提方法在 Concordia 大学的 CENPARMI 手写体数字数据库上的实验结果表明:利用这些特征能有效区分相似字符,提高了手写数字的识别率.

2 基于主曲线的手写数字特征提取

在众多的数字识别方法中,按使用特征的不同可分为基于结构特征的方法[8]和基于统计特征的方法. 对如何有效地提取图形的结构特征,研究者们也提出了许多方法[7,9,10],但如何选择一种方法使它能更好地提取出反映图像信息的特征是模式识别过程中重要的一步. 由于主曲线对数据的信息保持性好,它能真实地反映数据的形态,即曲线是数据集合的“骨架”,而非“脊梁骨”,因此它能很好地反映数据分布的结构特征,而且从主曲线定义可知,它与中轴线(与轮廓线等距离的穿过图形的光滑曲线)定义有明显的相似性. 鉴于主曲线的这些优点,本文选用它来作为手写数字的结构特征,并把它与常规细化方法进行了比较.

2.1 主曲线

定义 1. 主曲线. 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

$$f(s) = E(Y \mid s_f(y) = s),$$

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,它能真实地反映数据的形态. 图 1 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,它较好地勾画出了原始信息的轮廓.

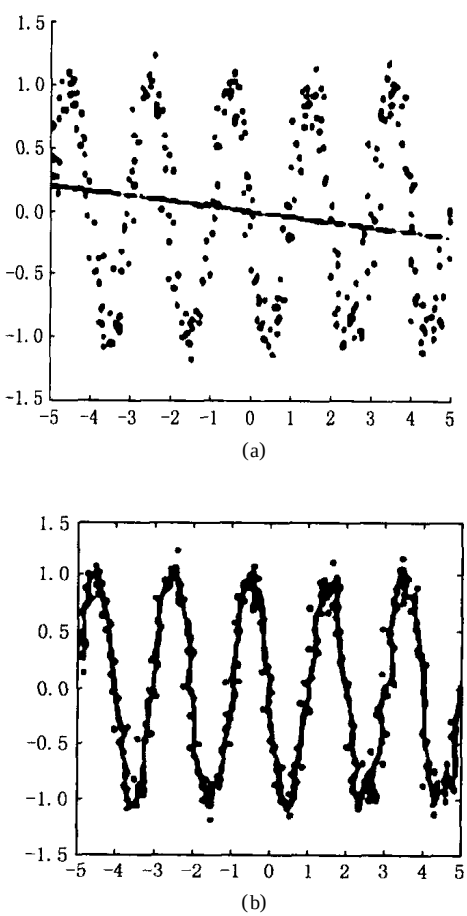


Fig. 1 The contrast graph between the first principal component line and principal curve. (a) First principal components and (b) Principal curve.

图 1 数据第一主成分与主曲线的对照图. (a) 第一主成分线;(b) 主曲线

2.2 特征提取

我们采用推广的多边形(PL)主曲线算法来提取手写数字的骨架结构. 算法主要有以下几步组成:

①初始化步. 它采用传统的 Suzaki-Abe 算法来获取原始字迹模板的近似骨架图 G_{vs}. G_{vs}由两个集合 V, S 构成,其中 $V=\{v_1, v_2, \cdots, v_n\} \subset R^d$ 是顶点的集合, $S=\{(v_{i1}, v_{j1}), \cdots, (v_{ik}, v_{jk})\}=\{s_{i1j1}, \cdots, s_{ikjk}\}$是边的集合.

②拟合-光滑步. 这一步的目的是调整骨架图 G_{vs}的光滑性,使之更好地拟合字形. 对给定数据集 $X_n=\{x_1, x_2, \cdots, x_n\}$,用距离惩罚函数 $E(G)=\Delta(G)+\lambda P(G)$ 取最小来优化骨架图,其中 $\Delta(G)=\frac{1}{n}\sum_{i=1}^{n}\Delta(x_i f)$ 是表示数据集中点到图形 G 中曲线的距离平方的平均值, $P(G)=\frac{1}{k+1}\sum_{i}^{k+1}P_v(v_i)$ 是关于 G 中曲线的平均曲率惩罚函数. $\Delta(G)$ 值较小表示骨架图可以较好地拟合数据, $P(G)$ 值较小能保证骨架图的光滑性较好. 在这一步中首先执行投影步,在投影步中把数据集 X_n 划分到属于骨架图顶点和边的最近邻区;其次做顶点优化步,即调整骨架图 G 中顶点和边的位置,使得距离惩罚函数 $E(G)$ 取得局部最小值.

③重构步. 这一步是对拟合-光滑步的补充完善. 在重构步中利用骨架图的几何性质对顶点和边的结构进行修改,消除或校正初始图形的瑕疵. 例如删除短分支、删除小圈等.

用推广的 PL 主曲线算法提取手写数字的骨架图见图 2.

Fig. 2 Digit skeleton graphs after the behavior of the generalized PL principal curve algorithm.

图 2 用推广的 PL 主曲线算法提取数字骨架

用推广的 PL 主曲线算法提取如图 2 所示的手写数字的结构特征后,我们可知每一个数字由几条光滑的曲线组成. 一条曲线是由一系列点 $p_{i_1,\cdots,i_l}=(v_{i_1}, \cdots, v_{i_l})$ 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 1 个数字由 3 条光滑曲线组成.

2.3 与细化方法比较

由于主曲线与细化方法都是用来提取字符的骨架,所以在这一部分,我们从下面几方面把它与细化方法进行比较.

(1) 在基于细化的特征提取算法中,由于字符图像信息是以位图形式存储的,且体现的是像素点信息,因此在特征选取过程中有诸多不便,需要对整幅图像进行扫描等处理. 而使用主曲线就完全不同了,主曲线所得到的结果是以矢量形式存储的,这样就有利于特征选取.

(2) 由于经过了拟合光滑步和重构步,使处理后的骨架更能准确地反映字符的整体拓扑结构,因此,在其基础上选取的字符特征比原来基于细化而得到的特征不但数量少而且要更加准确. 这将对提高字符识别率有很大帮助.

(3) 采用主曲线符合人们对字符是按曲线理解而非像素点的特点,即,从图论的角度看,它更有利于进行有效特征的选取.

(4) 由于在提取主曲线时,需要对大量数据进行投影,所以其时间效率比细化差一点.

(5) 主曲线是字符识别的一种新的尝试,有着更大的发展空间和前景.

3 特征的预处理

由于手写数字的书写因人而异,因时而变,形态变化十分巨大. 为了更有利于特征选取,我们在用推广的 PL 算法提取完主曲线特征后,在进行特征选取前先对近似回路和回路外的短分支进行处理. 例如把  等处理成 . 把 处理成 . 具体处理算法如下:

(1) 对一条曲线而言,如果存在一点 v_i 到其中某一端点 v_j 的距离 $d(v_i, v_j)$ 除以这两点间曲线的长度 $\sum_{k=i}^{j-1}d(v_k, v_{k+1})$ 小于某一参数,即

$$\frac{d(v_i, v_j)}{\sum_{k=i}^{j-1}d(v_k, v_{k+1})} < parameter1,$$

则认为这两点间形成一条近似回路. 我们添加一条线段 (v_i, v_j) 使之构成近似回路. 经大量训练样本训练可知参数 $parameter1=0.213$ 时效果最好.

(2) 对两条曲线而言,例如一条曲线的两端点为 v_i 和 v_k,另一条曲线的端点为 v_k 和 v_j. 则 v_i 和

v_j 的距离 $d(v_i, v_j)$ 除以这两条曲线长度之和 $\sum_{p=1}^{k-1} d(v_p, v_{p+1}) + \sum_{q=k}^{j-1} d(v_q, v_{q+1})$ 小于某一参数,即

$$\frac{d(v_i, v_j)}{\sum_{p=1}^{k-1} d(v_p, v_{p+1}) + \sum_{q=k}^{j-1} d(v_q, v_{q+1})} < parameter2,$$

则认为这两条曲线形成一条近似回路. 我们添加一条线段 (v_i, v_j) 使之构成近似回路. 经大量训练样本训练可知参数 $parameter2 = 0.267$ 时效果最好. 数字中近似回路的个数称为近似回路数.

(3) 设曲线的端点为 v_i 和 v_j,回路的端点为 v_j. 如果曲线的长度 $\sum_{k=1}^{j-1} d(v_k, v_{k+1})$ 除以回路的长度 $\sum_{k=1}^{j-1} d(v_k, v_{k+1})$ 小于某一参数,即

$$\frac{\sum_{k=1}^{j-1} d(v_k, v_{k+1})}{\sum_{k=j}^{j-1} d(v_k, v_{k+1})} < parameter4,$$

则认为曲线相对于回路可删除,否则就不可删除. 经大量训练样本训练可知参数 $parameter4 = 0.158$ 时效果最好.

4 数字特征选取

4.1 特征选择

提取数字的主曲线并预处理后,如何充分有效地利用数字的结构信息来选取出最能反映分类本质的特征是模式识别过程中关键的一步,即如何进行特征选取对提高识别率是非常重要的. 虽然手写数字的书写因人而异、因时而变,形态变化十分巨大,但用主曲线来作为其结构特征受字体变形的影响小. 通过对数字主曲线特征详细分析后发现:回路数是区分数字的重要特征,且易于检测. 利用回路数这一整体结构特征可先对数字进行粗分类,可分 3 类:

{0 2 3 4 6 9}, {8}, {1 2 3 4 5 7 7}.

对数字粗分类后,我们再根据笔画数、凸曲线数、凹曲线数、凸凹点及交点的相对位置、笔画端点相对于回路的位置等细节特征将数字进一步区分. 例如,我们通过分析发现不管 3 如何书写其凸曲线数必为 2,其他数字则不存在这个特点;8 不管如何写预处理之后其回路数必为 2;端点与回路的相对位置可区分 6 和 9 等规则.

4.2 计算回路数

由于数字是由笔画本身连接或交叉而成,因此用主曲线进行特征提取后得到的图像可看做是一幅连通图.

这里设连通图 G 中边的数量为 E,顶点为 v,顶点数是 n,每个顶点的度数为 $deg(v)$. 如果以端点、二叉点、三叉点和四叉点作为连通图的顶点,那么显然,其顶点度数分别与它们的连通分量相同,即分别为 1,2,3 和 4. 为了检测连通图中的闭合曲线,即回路数,这里结合离散数学中图论知识可给出下面两个定理:

定理 1. $\sum_{v \in V(G)} deg(v) = 2E$,其中,$V(G)$ 为 G 中顶点的集合.

定理 2. 设 G 是连通图,则 G 中有闭合曲线的充要条件是 $E \geq n$,且当 $E = n$ 时,G 中有且只有一条闭合曲线,当 $E > n$ 时,G 中必有多条闭合曲线. (证明略).

下面我们给出数字图像中是否存在闭合曲线的检测方法:

(1) 在主曲线图中抽取端点、二叉点、三叉点和四叉点,分别记录其个数 n_1, n_2, n_3 和 n_4.

(2) 用 Euler 公式计算出图中的边数 $E = (n_1 + 2n_2 + 3n_3 + 4n_4)/2$;顶点数 $n = n_1 + n_2 + n_3 + n_4$.

(3) 若 $E > n$,则图中至少有 2 条闭合曲线存在;

若 $E = n$,则图中有且仅有 1 条闭合曲线

若 $E < n$,则图中没有闭合曲线存在.

4.3 细节特征的选取

提取数字的主曲线后,通过详细分析,我们发现除回路数这一整体特征外,笔画数、凸曲线数、凹曲线数、凸凹点及交点的相对位置、笔画端点相对于回路的位置、是否为直线是区分数字的重要细节特征.

下面具体说明有关的细节特征:

数字中光滑曲线段的个数称之为笔画数(在这里一个回路表示一画),容易知道,笔画数 = 最大分叉数 - 回路数. 例如  有一个四叉点,因而最大分叉数是 4,又因为回路数为 1,所以笔画数为 3.

除回路以外,组成数字的各曲线中,如果曲线在 X 方向具有极大值,则称之为凸曲线. 凸曲线的条数称为凸曲线数. 例如  的凸曲线数为 2. 如果曲线在 X 方向具有极小值,则称之为凹曲线. 凹曲线的条数称为凹曲线数. 例如  的凹曲线数为 1.

在这里定义的直线是近似直线. 设曲线的两个

端点为 v_i 和 v_j,如果曲线满足两端点间距离 $d(v_i, v_j)$ 除以曲线长度 $\sum_{k=i}^{j-1} d(v_k, v_{k+1})$ 大于某一参数,即

$$\frac{d(v_i, v_j)}{\sum_{k=i}^{j-1} d(v_k, v_{k+1})} > parameter3,$$

则认为这条曲线是直线,否则称曲线不是直线. 经大量训练样本训练可知参数 $parameter3 = 0.847$ 时效果最好.

5 实验结果与分析

对新样本的分类过程如下:对训练集中每个样本,首先抽取其主曲线,然后利用曲线的回路特征进行粗分类,最后利用细节特征进行细分类. 在这里我们把基于主曲线的分类器命名为 CBPC.

本实验在 Concordia 大学的 CENPARMI 手写体数字数据库上进行,该标准数据库有 4000 个训练集,2000 个测试集. 实验结果表明了利用这些特征能有效区分相似字符,提高了手写数字的识别率. 识别结果如表 1 所示:

Table 1 Recognition Results on Different Classifier
表 1 不同分类器上的识别结果

Classifier	Error Rate (%)
MQDF[11]	4.8
S. W. Lee[11,12]	2.20
Local Learning Framework[13]	1.90
Virtual SVM[14]	1.30
SVC-rbf[15]	1.10
CBPC	0.923

上述实验从 CENPARMI 标准数据库出发来比较本实验和常用的几个手写数字识别器的性能. 实验数据充分表明了本文提出的基于主曲线的特征选取方法使得分类器整体性能有了新的提高.

6 结论和进一步研究工作

本文提出了一种新的基于主曲线的手写数字特征选取方法. 从理论上讲,训练集的规模越大越好,但从实验中发现训练集规模达到一定程度时,本算法的识别率已很高且稳定. 初步观察发现它所要求的训练集规模比其他方法相对要小,这样就会节省不少训练时间. 下一步将对本算法的时间和空间复杂度进行深入的理论分析与实验验证,并与其他方法进行分析比较. 同时为了进一步完善本系统的性能,下一步我们将考虑如何把数字的统计信息引入到本方法中,充分利用结构信息与统计信息来设计更好的分类器.

参 考 文 献

1 T. Hastie. Principal curves and surfaces. Laboratory for Computational Statistics, Stanford University, Department of Statistics, Technical Report: 11, 1984

2 Zhang Junping, Wang Jue. An overview of principal curves. Chinese Journal of Computers, 2003, 26(2): 129～146 (in Chinese)
(张军平, 王 珏. 主曲线综述. 计算机学报, 2003, 26(2): 129～146)

3 J. D. Banfield, A. E. Raftery. 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

4 B. Kegl, A. Krzyzak, *et al*. A polygonal line algorithm for constructing principal curves. In: Proc. Neural Information Processing System. Denver Colorado, USA: Computer Press, 1999. 501～507

5 J. J. Verbeek, N. Vlassis, B. Krose. A *k*-segments algorithm for finding principal curve. Computer Science of Institute, University of Amsterdam, Technical Report: IAS-UVA-00-11, 2000

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

7 B. Kegl, A. Krzyzak, *et al*. Piecewise linear skeletonization using principal curves. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002, 24(1): 59～74

8 Lou Zhen, Hu Zhongshan, Yang Jinyu. Handwritten Arabic digit recognition based on segment features of skeleton. Chinese Journal of Computers, 1999, 22(10): 1065～1073 (in Chinese)
(娄震, 胡钟山, 杨静宇. 基于轮廓分段特征的手写体阿拉伯数字识别. 计算机学报, 1999, 22(10): 1065～1073)

9 Liu Junyi, Wang Runsheng. Skeletonization algorithm for gray-scale patterns based on erosion simulation. Acta Electronica Sinica, 2001, 29(9): 1259～1262 (in Chinese)
(刘俊义, 王润生. 基于冲刷模拟的灰度模式骨架化算法. 电子学报, 2001, 29(9): 1259～1262)

10 S. W. LEE, L. Lam, C. Y. Suen. A systematic evaluation of skeletonization algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 1993, 7(5): 1203～1225

11 L Y. Lecun, L. Jackel, L. Bottou, *et al*. Comparison of learning algorithms for handwritten digit recognition. In: F. Fogelman, P. Gallinari, eds. Int'l Conf. Artificial Neural Networks. Paris: AI Computer Press, 1995. 53～60

12 W. Lee. Multilayer cluster neural network for totally unconstrained handwritten numeral recognition. Neural Networks,

1995, 8(5): 783～792

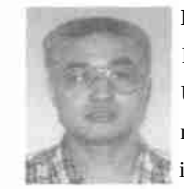
13 J. X. Dong, A. Krzyzak, C. Y. Suen. Local learning framework for handwritten character recognition. Engineering Applications of Artificial Intelligence, 2002, 15(2): 151～159

14 Bailing Zhang, Minyue Fu, Hong Yan, *et al*. Handwritten digit recognition by adaptive-subspace self-organizing map. IEEE Trans. Neural Network, 1999, 10(4): 589～603

15 L. N. Teow, K. F. Loe. Robust vision-based features and classification schemes for off-line handwritten digit recognition. Pattern Recognition, 2002, 35(11): 2355～2364

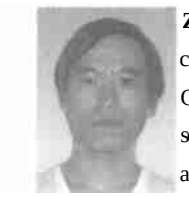
Zhong Hongyun, born in 1972. Ph. D. candidate in Tongji University, Shanghai, China. Her current research interests include pattern recognition, date mining and principal curve.

张红云,1972 年生,博士,主要研究方向为模式识别、数据挖掘、主曲线.



Miao Duoqian, born in 1964. Professor and Ph. D. candidate supervisor in Tongji University, Shanghai, China. His current research interests include artificial intelligence, pattern recognition, date mining, rough set and principal curve.

苗夺谦,1964 年生,教授,博士生导师,主要研究方向为人工智能、模式识别、数据挖掘、粗糙集理论、主曲线.



Zhang Dongxing, born in 1980. Master candidate in Tongji University, Shanghai, China. His current research interests include software and theory of computer, rough set and principal curve.

张东星,1980 年生,硕士研究生,主要研究方向为计算机软件与理论、粗糙集理论、主曲线.

Research Background

This research work is supported by the National Nature Science Foundation of China, the National 973 Program No. 2003CB316902, and the Shanghai S&T Committee Important Science and Technology Program.

Extraction and choice of features are critical to improve the recognition rate of handwritten digits. It is highly noted by the research area, and researchers have put forward many methods of feature extraction. Comparing with these methods, principal curves can preferably reflect the patter features of data. Principal curves are smooth self-consistent curves that pass through the "middle" of the distribution. During digit feature selection, firstly principal curves are used to extract the structural features of training data; Secondly the classification features used for digits coarse classification and precise classification are chosen by analyzing the structural features of principal curves in detail; Finally coarse classification and precise classification are separately carried out in the handwritten digits recognition. The Concordia University CENPARMI handwritten digit database is used in the experiment. The result of experiment shows that these features have good discriminating power of similar digits. The proposed method can effectively improve the recognition rate of off-line handwritten digits and provide a new approach to the research for off-line handwritten digits recognition.