ICS 35.240.30
L 67

中华人民共和国国家标准

GB/T 18790—2002

联机手写汉字识别技术要求与测试规程

Requirements and test procedure of on-line handwriting Chinese ideogram recognition

2002-07-18 发布　　2002-12-01 实施

中华人民共和国
国家质量监督检验检疫总局　发布

前　　言

本标准规定了联机手写汉字识别系统的汉字识别技术要求、测试规程。该标准的制定和实施将规范联机手写汉字识别系统的研究、开发和应用。

本标准的附录 A 和附录 B 是标准的附录，附录 C 是提示的附录。

本标准由中华人民共和国信息产业部提出。

本标准由中国电子技术标准化研究所归口。

本标准起草单位：中国电子技术标准化研究所、北京汉王科技有限公司、清华大学电子工程系。

本标准主要起草人：刘迎建、王立建、张立清、刘长松、钮兴昱、王宝艾。

中华人民共和国国家标准

联机手写汉字识别技术要求与测试规程

GB/T 18790—2002

Requirements and test procedure of on-line handwriting Chinese ideogram recognition

1 范围

1.1 本标准规定了联机手写汉字识别系统的汉字识别技术要求和测试规程。

1.2 本标准适用于微型计算机、手持式信息处理设备和数字化电器配置的联机手写汉字识别系统。

2 引用标准

下列标准所包含的条文，通过在本标准中引用而构成为本标准的条文。本标准出版时，所有版本均为有效。所有标准都会被修订，使用本标准的各方应探讨使用下列标准最新版本的可能性。

GB 2312—1980 信息交换用汉字编码字符集 基本集

GB 12345—1990 信息交换用汉字编码字符集 辅助集

GB 13000.1—1993 信息技术 通用多八位编码字符集(UCS) 第1部分：体系结构 (idt ISO/IEC 10646:1993)

GB 18030—2000 信息技术 信息交换用汉字编码字符集 基本集的扩充

3 定义

本标准采用下列定义。

3.1 **联机手写数据采集设备** on-line handwriting data capture device

是指鼠标、手写板、触摸屏等具备实时地将人们书写的汉字及字符轨迹转换成坐标点序列、形成电子数据的设备。

3.2 **联机手写汉字识别系统** on-line handwriting Chinese ideogram recognition system

是指使用微型计算机或具有计算能力的系统，对联机手写电子数据采集设备采集的手写电子数据进行处理与辨识，获得相应的标准内码的系统。

3.3 **样本** sample

一个完整的、由联机手写电子数据采集设备采集的、符合第4章中的汉字或字符的电子数据称为一个样本。

3.4 **样本文件** sample file

是指由多个样本按附录B数据格式组成的文件，称为样本文件。

3.5 **样本库** library of sample file

多个同类型样本文件组成的文件集合称为样本库。

4 识别字符集的范围

联机手写汉字识别系统识别的最小字符集应是GB 2312中全部汉字字符(包括偏旁部首)以及附录A中的非汉字字符。字符扩展时，联机手写汉字识别系统应识别GB 12345，或GB 18030，或GB

中华人民共和国国家质量监督检验检疫总局 2002-07-18 批准　　2002-12-01 实施

13000.1 字符集中的全部汉字。

5 识别技术要求

5.1 识别率

5.1.1 对工整样本库的识别要求

a) 对 GB 2312 中所有汉字的识别率应大于 94%；

b) 对附录 A 中的非汉字字符的识别率应大于 80%；

c) 若厂商声明支持 GB 12345，或 GB 13000.1，或 GB 18030 字符集，则以上字符集的识别率应大于 85%。

d) 对任何字符集而言，单字识别率应大于 50%。

5.1.2 对乱笔顺样本库的识别要求

对乱笔顺样本库，识别率应大于 60%。

5.2 识别速度

在测试软件运行的平台上，识别速度应优于 1.5 s/字。

6 测试规程

6.1 标准测试样本库的建立

a) 由信息处理产品标准符合性检测中心分别用压力式手写板、电磁感应手写板等设备各采集由一定数量的人、在工整书写提示下自然书写的汉字样本，经整理后，建成标准测试样本库；

b) 标准测试样本库中包括两部分样本，其中一部分是工整书写的样本，称为工整样本库；少部分是人工方式处理的、打乱了笔划顺序的工整样本，称为乱笔顺样本库。

6.2 由信息处理产品标准符合性检测中心提供标准测试样本文件的数据格式及结果文件格式，并提供至少一个样本文件，供参测单位调试测试程序。

6.3 参测单位向标准符合性测试机构提交联机手写汉字识别技术的测试软件，由信息处理产品标准符合性检测中心测试，并提供测试结果。

非微型计算机平台上的参测单位，除测试软件外，还应提供能完成测试工作的运行平台。

6.4 测试结果计算

6.4.1 识别速度

识别速度测试结果按下式计算：

$$识别速度 = T/N$$

式中：N——测试样本库中样本总数；

T——识别系统从开始读取测试数据至将识别结果记录到媒体上所用的时间。

6.4.2 识别率

识别率测试结果按下式计算：

$$识别率 = CN/NN$$

式中：NN——样本库中样本总数；

CN——经统计第一选识别结果正确的样本数。

6.4.3 单字识别率

单字识别率测试结果按下式计算：

$$单字识别率 = CS/NS$$

式中：NS——样本库中某一汉字的样本总数；

CS——经统计第一选识别结果正确的该汉字样本数。

附 录 A
（标准的附录）
联机手写汉字识别系统的识别字符集的非汉字字符集

联机手写汉字识别系统至少识别如下非汉字字符：

A1 数字：

0 1 2 3 4 5 6 7 8 9

A2 大写英文字符：

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A3 小写英文字符：

a b c d e f g h i j k l m n o p q r s t u v w x y z

A4 标点符号：

！″，：；？、。′（）·—…〈〉《》# $ % &. * ＋－．/＜＝＞@ ¥［＼］^ _ '｛｜｝～ £

A5 手势（箭头代表书写方向，括号内为该手势的内码）：

空格（0x0020）：

回车（0x000D）：

回删（0x0008）：

删除（0x001e）：

附 录 B
（标准的附录）
标准测试样本文件格式

本附录规定了标准测试样本库中的样本文件的扩展名和文件格式。本附录对于提供和使用联机手写汉字识别系统的各方提出共同遵守的约定。

B1 标准测试样本文件的扩展名

标准测试样本文件的扩展名统一为“.POT”。

B2 标准测试样本文件格式

文件格式：标准测试样本文件中存放的是连续的 POT 数据块；无其他额外的格式信息。

POT 数据块定义：每个 POT 数据块中包含一个手写汉字的字型采样数据和内码等信息，详见下表。

位移	内容
0 WORD	数据块长度，以字节（BYTE）为单位，假设为 n
2 DWORD	本数据块所代表的汉字内码，例如“啊”＝0x0000b0a1
6 WORD	手写样本的总笔划数，假设为 N

8 WORD	第一笔的第一个点的 X 方向坐标值
10 WORD	第一笔的第一个点的 Y 方向坐标值
…	
i WORD	0xFFFF，第一笔的结束标志
$i+1$ WORD	0x0000，第一笔的结束标志
…	
$n-7$ WORD	0xFFFF，第 N 笔的结束标志
$n-5$ WORD	0x0000，第 N 笔的结束标志
$n-3$ WORD	0xFFFF，本字的结束标志
$n-1$ WORD	0xFFFF，本字的结束标志

其中：

1 WORD＝2 BYTE（低字节在前，高字节在后），例如：0x1234，在文件中的字节排列顺序为 34 H，12 H；

1 DWORD＝2 WORD（低字在前，高字在后），例如：0x12345678，在文件中的字节排列顺序为 78 H，56 H，34 H，12 H。

附 录 C
（提示的附录）
联机手写汉字识别系统程序接口规范

本附录定义了基于标准 C 语言的联机手写汉字识别程序接口规范。本附录对于提供和使用联机手写汉字识别系统的各方提出共同遵守的约定。

C1 API 文本细则

全部接口函数共 10 个，分列如下：

C1.1 char * OLGetBrand(void)

说明：获得识别程序的提供厂商和版本说明。

返回值：

成功，返回一个字符串指针，最多 1024 个字符；

否则，返回 0。

C1.2 DWORD OLGetVersion(void)

说明：获取识别程序版本号。

返回值：

成功，返回识别程序版本号，高字为主版本号，低字为子版本号；

否则，返回 0。

C1.3 char * OLGet Date(void)

说明：获取识别程序的提交时间。

返回值：

成功，返回一个字符串指针，其中时间以“yyyy-mm-dd”形式提供；

否则，返回 0。

C1.4 int OLInit(void)

说明：本函数用来初始化识别程序，装入识别字典。

返回值：

成功，返回一个非零值；

否则，返回 0。

C1.5 int OLClose(void)

说明：本函数用来释放识别字典。

返回值：

成功，返回非零；

否则，返回 0。

参见：OLInit

C1.6 DWORD OLSetRange(DWORD range)

说明：本函数用来设置识别字符集的范围。

参数：

range：指定的识别范围。定义如下：

bit0：小写英文

bit1：大写英文

bit2：数字

bit3：常用标点 8 个，包括：，。、?!″：；

bit4：扩展标点，包括：′（）…〈·〉《》—

bit5：常用符号，包括：£￥#.＝/＞$－%＋＜*@&.

bit6：扩展符号，包括：～{}^ '\]_[|

bit7：手势 4 个，包括：空格(0x0020)、回车(0x000d)、回删(0x0008)、删除(0x0010)

bit8：偏旁部首

bit9：GB 2312 一级国标简体汉字

bit10：GB 2312 二级国标简体汉字

bit11：GBK 3 区中的汉字

bit12：GBK 4 区中的汉字

bit13 以上：保留

返回值：

成功，返回旧的识别范围；

否则，返回 0。

参见：OLRecognize，OLGetRange

C1.7 DWORD OLGetRange(void)

说明：本函数用来读取识别字符集的范围。

返回值：

成功，返回当前的识别范围，参见 OLSetRange；

否则，返回 0。

参见：OLSetRange

C1.8 int OLSetCandidateNum(int num)

说明：本函数用来设置识别候选字的数量。

参数：

num：设置识别候选字的数量，缺省值为 10，最大值为 20。

返回值：

成功，返回旧的候选字的数量；

否则，返回 0。

参见:OLGetCandidateNum,OLRecognize

C1.9 int OLGetCandidateNum(void)

说明:本函数用来读取识别候选字的数量。

返回值:

成功,返回当前的候选字的数量,参见 OLSetCandidateNum;

否则,返回 0。

参见:OLSetCandidateNum

C1.10 int OLRecognize(WORD * lpTrace,WORD * lpResult)

说明:本函数用来识别输入的笔迹。

参数:

lpTrace:输入的笔迹数据指针,其空间由应用程序申请,数据类型为 WORD(2 byte),格式如下:

(x0,y0)(x1,y1)...(0xffff,0)...(0xffff,0)...(0xffff,0xffff)

^ 笔划结束标志 ^ 字结束标志

lpResult:存放识别结果的数据指针,其空间由应用程序申请,不应少于由 OLSetCandidateNum 设字的候选字个数 * 6;识别程序将识别结果及可信度得分填入其中。

每个结果占 2～4 个字节(参见 GB 18030);半角字符和手势结果均为两字节,其高位均为 0。全部识别结果之后是每个结果的可信度得分,每个得分表示为 1 个 WORD,按照识别结果排列的顺序排列。例如,共有 4 个识别结果:"啊阿可何",可信度得分分别为:100,90,80,70;那么,lpResult 中的排列顺序就是:啊,阿,可,何,100,90,80,70。

返回值:

成功,返回识别结果的个数;

拒识,返回 0。

参见:OLSetRange,OLSetCandidateNum

备注:

1 WORD=2 BYTE;可定义为 #define WORD unsigned short int。

1 DWORD=2 WORD。可定义为 #define DWORD unsigned int(32 位系统)或 #define DWORD unsigned long int(16 位系统)。

中 华 人 民 共 和 国
国 家 标 准
联机手写汉字识别技术要求与测试规程
GB/T 18790—2002
*
中国标准出版社出版
北京复兴门外三里河北街 16 号
邮政编码:100045
电话:68523946 68517548
中国标准出版社秦皇岛印刷厂印刷
新华书店北京发行所发行 各地新华书店经售
*
开本 880×1230 1/16 印张 3/4 字数 14 千字
2002 年 12 月第一版 2002 年 12 月第一次印刷
印数 1—1 500
*
书号:155066·1-19086 定价 10.00 元
网址 www.bzcbs.com
*
科 目 631—474

GB/T 18790-2002