



扫描仪图文扫描

吴越编著
北京图书馆出版社

青少年电脑实用知识丛书

扫描仪图文扫描

吴 越 编著

北京图书馆出版社

图书在版编目(CIP)数据

青少年电脑实用知识丛书 / 吴越编著. — 北京: 北京图书馆出版社, 1998. 12

ISBN7-5013-1540-X

I. 青… II. 吴… III. 电子计算机—青少年读物 IV. TP3

中国版本图书馆 CIP 数据核字(98)第 30818 号

书 名 青少年电脑实用知识丛书
著 者 吴 越 编著

出版 北京图书馆出版社(原书目文献出版社)
发行 (100034 北京西城区文津街 7 号)
经销 新华书店
印刷 湖南广播电视报印刷厂

开本 850×1168(毫米) 1/32
印张 53
字数 1200(千字)
版次 1998 年 11 月第 1 版 1998 年 11 月第 1 次印刷
印数 1—4000(套)

书号 ISBN 7-5013-1540-X/G·417
定价 70.00 元(全套 10 册)

总 序

我是大陆最早使用电脑从事文学创作并处理日常事务的作家之一。作为“先行者”，宣传、推广电脑的优越性并辅导“后来人”，当然是责无旁贷的事情。为此，今年年初我应北京图书馆出版社之约，将我使用和教学电脑的经验写成了十本一套的《青少年电脑知识普及丛书》。

那套丛书，以文字处理为中心，先讲解电脑的基本知识，然后介绍目前中国比较通用又好学好用的几种汉字输入法和屏幕编辑法，旁及电脑病毒的防治、工具软件的使用、简单的维修方法，再介绍几个有趣的电脑游戏，作为调剂。一般说来，那套丛书作为“入门”的教材，已经够了。

那套丛书出版以后，我收到了全国各地许多读者写来的信和打来的长途电话，反映他们在学习电脑的过程中所遇到的疑难问题并提出各种各样的要求，希望我提供帮助。

这些意见和要求中，除了软件问题之外，反映得最多的，无非是“如何开发电脑的其他功能”这个话题。有个读者还说了一句俏皮话：“您带着那么多人一下子涌进了电脑这个大门，总不能叫我们都站在院子里呀？总也得让有些人再走几步，来一个‘登堂入室’吧？”——也就是说，在这些初学电脑的人中间，似乎出现了一个“断层”。

这似乎应了一句古话：“师傅领进门，修行在个人。”再想提高，师傅就不管了。

电脑是高科技产品，是尖端科学，功能十分强大，各行各业各个领域的人都可以用它来为自己服务。用电脑来处理汉字，其实只是电脑功能中最基本的也是最小的那一部分。

于是就产生了这样一个矛盾：家里放台电脑，除了打字之外，没有其他用处，未免“大材小用”，浪费了设备；想再学点儿别的用途，又怕花费时间精力太多，得不偿失。

其实，如果仅仅从“家用”两字着眼进行开发，有一些项目，例如用它来管理通信录、计算银行储蓄利息、为多家

住户结算一个月的水费电费、让电脑成为孩子的家庭教师、用电脑来学习英语或进行简单的绘图、作曲等等，是不需要太多的专业知识，也不用花费很多的时间，就可以掌握的。此外，会计电算是目前的热门，扫描仪的使用也日见普遍，如果打算当一个专职的文秘人员，这些都是基本功。再说，Internet 已经像一个魔术师似的把全世界的电脑联接起来，年轻人大多愿意自投罗网。只要您的电脑档次不是太低，每月一二百元的电话费也出得起，足不出户就能到全世界去漫游，广交朋友，未始不是一件既有趣又有益的事情。

中国目前电脑使用的情况，一方面是大量的电脑拥有者虽然经常和电脑打交道，却长期停留在初学者的水平上得不到提高；另一方面，则是从事电脑工程的专业人员，他们以电脑为业，尽管道行有高低，水平分上下，但总是已经“登堂入室”的“此中人也”。一个初级程序员所知道的东西，无疑应该比进了大门还“站在院子里”的人要高得多。

这两部分人中间的空档，就是所谓的“断层”。一个刚刚走进院子的“门外汉”，想要从院子里再走进厅堂中，是需要一个台阶的。这个台阶，可以是各种各样的电脑学习班，也可以是形形色色的电脑书报和杂志。

但是，要请电脑专家来写这样的书，却也有些困难：有些人怕“跌份儿”，不屑于写这种通俗的东西；有些人则又写惯了学术论文，一动笔就带三分学院气、七分书生气，写不来这种给“科盲”看的通俗文章。

这套丛书，就是专为适应这部分读者的需求而写的。我是个通俗文学作家，也是个科普读物作家。“通俗化”不但我的终身事业，也是我的毕生追求。我写这套书，应该说同时也就是我自己总结、提高的过程。因此，这套丛书正是我自己从一个台阶迈上另一个台阶的脚印。

但愿这套丛书能成为您“登堂入室”的台阶。

吴 越

1998年10月1日于北京

目 录

导 言	1
第 一 章 图文扫描与字符识别	4
一、硬件设备：扫描仪	5
(一)UNISCAN 系列扫描仪的种类	6
(二)UNISCAN 的安装	7
二、OCR 技术在我国的发展概况	10
三、清华紫光 OCR 6.0 版介绍	13
(一)性能和指标	13
(二)软件的安装	15
(三)文字扫描与识别	17
(1)界面介绍(17)	
(2)菜单操作(19)	
(3)各项有关设置(19)	
(4)开始扫描(22)	
1)关于扫描模式的选择(22)	
2)关于扫描分辨率的选择(23)	
3)功能按钮(23)	
4)预扫描(24)	
5)正式扫描(25)	
(5)图像处理(27)	
1)自动纠斜(27)	
2)定点切块(27)	
3)块删除(31)	
4)块清除(31)	
5)横/竖排版(31)	
6)图像信息(32)	
7)表格线线宽设置(32)	

(6)图像系统(32)	
(7)文字识别(37)	
(8)后处理(38)	
(四)清华 OCR 6.0 普及版简介	42
四、清华紫光 OCR 7.0 版介绍	44
(一)系统运行环境	44
(二)软硬件的安装	44
(三)启动及设置	45
(四)扫描文件操作	48
(五)清华紫光 OCR 7.0 版各项功能介绍	53
A. 文件栏	53
(1)扫描设置(53)	
(2)另存为(57)	
(3)导出(58)	
B. 编辑栏	60
(1)剪切板功能(61)	
(2)改错字功能(61)	
1) 系统自校可疑字(61)	
2) 用户校核错别字(62)	
C. 图像栏	63
(1)旋转图像(63)	
(2)反转图像(65)	
(3)剪裁图像(65)	
(4)清除区域(66)	
(5)反转区域(66)	
(6)删除区域(66)	
(7)删除所有区域(66)	
(8)删除框线(67)	
(9)删除所有框线(67)	
(10)恢复图像(67)	
(11)设置版面属性(67)	
D. 命令栏	72

(1)版面分析(72)	
1) 自动版面分析	
2) 手动版面分析	
(2)检测表格线(73)	
1) 框线检测(73)	
2) 框线修补(74)	
(3)识别(74)	
(4)识别所有页(77)	
(5)逆行序(77)	
(6)用户字学习(78)	
(7)修改用户库(78)	
(8)发送文本(79)	
E. 显示栏.....	79
(1)局部图像显示的比例(79)	
1) 放大(80)	
2) 缩小(80)	
3) 选择比例(80)	
(2)显示图像(80)	
(3)页面选择(81)	
(4)工具条(81)	
(5)状态行(81)	
(六)清华紫光 OCR 7.0 普及版介绍	81
(七)清华文通 OCR 标准版介绍.....	84
第二章 图像的扫描与编辑	86
(一)“扫描大师”的功能	86
(二)“扫描大师”的安装	87
(三)启动“扫描大师”	88
(四)用“扫描大师”扫描图像.....	89
(五)工具箱中的图标按钮.....	91
(1)定义框(92)	
(2)加文字(92)	

- (3) 旋转(93)
- (4) 游动(94)
- (5) 画笔(94)
- (6) 吸管(94)
- (7) 画直线(94)
- (8) 放大镜(95)
- (9) 空心矩形(95)
- (10) 实心矩形(95)
- (11) 空心圆(95)
- (12) 实心圆
- (13) 选择颜色(96)
- (14) 画笔风格(97)
- (六) 菜单命令97
 - (1) 文件操作(97)
 - 1) 打开文件(98)
 - 2) 扫描图像(98)
 - 3) 打印(98)
 - 4) 存盘(99)
 - (2) 图像的编辑处理(99)
 - 1) 亮度和对比度(99)
 - 2) 反色(100)
 - 3) 平滑(100)
 - 4) 锐化(100)
 - 5) 旋转(101)
 - 6) 镜像(101)
 - 7) 缩放(102)
 - 8) 转黑白图像(103)
 - 9) 转灰度图像(103)
 - 10) 转 256 色图像(103)
 - 11) 转真彩色图像(104)
 - 12) 还原(104)
 - 13) 剪切(104)

- 14)复制(104)
- 15)粘贴(105)
- 16)清除(105)
- (3)显示(106)
 - 1) 1:1 显示(106)
 - 2)适应窗口显示(106)
 - 3)缩放显示(106)
 - 4)图像信息(106)
 - 5)系统信息(107)
 - 6)显示/隐藏工具箱(107)
- (4)选项(107)
 - 1)手动/自动扫描(107)
 - 2)中/英文切换(107)
 - 3)改变主窗口风格(107)
- (5)窗口(108)
 - 1)层叠(108)
 - 2)平铺(108)
 - 3)安排图标(108)
 - 4)关闭所有文档窗口(108)
- (6)自动扫描(109)

导 言

汉字输入电脑，有许多方法。除了敲打键盘之外，还有语音输入法、手写输入法和扫描输入法等等。不通过键盘在电脑中输入汉字，因为不使用汉字编码，所以统称“无码输入法。”

严格地说，“无码输入法”，实际上也是有码的，不过用的是“内码”，是由电脑自动识别以后进行转换的，用户不必再记任何编码。因此无码录入，也叫自动录入。

既然无码录入可以不用记忆任何编码，那么为什么不大力推广，让大家都来使用，还要学那挺难学的键盘输入法干什么？

无码录入固然有它不用学习编码的长处，但也有它并非善尽美的短处。

例如语音录入，一个人对着话筒说话，电脑根据语音转换为汉字内码，记录下来的不是声音而是文字，从理论上说这简直太妙了；但在实际操作中问题还很多：

第一，语音录入对电脑的档次要求比较高。以 IBM 公司开发的语音系统为例，要求电脑的档次必需是 586/166 以上。尽管现在高档电脑的价格已经落得很低，586 电脑的价格比高档的 486 电脑贵不了多少，新买电脑的人，除了家庭经济十分拮据，只能买台低档电脑用来打打字之外，很少有人会去选择 486 以下的机器。但是目前全国 500 多万台电脑中，586/166 以上档次的究竟还是少数。再说，多数已经买了电脑的作家、编辑、教师等等，只要不是新近买的，绝大多数人用的还都是 486 以下。据我所知，相当多的知名作家，用的还是 386 甚至 286 机器。因为对于用键盘输入文字的人来说，如果不强调版面格式和字体变化，有一台 286 电脑，确实已经够用了的。

因此，为了实现语音输入而要求已经配备了电脑的人扔掉(一台有硬盘的单显 286 二手电脑目前只能卖四五百块钱)完全能用的低档电脑而去更换高档电脑，是很不现实的。

第二，语音输入文字的正确率目前还达不到要求。几千年来，汉字都是以“字”为单位书写，字与字之间的组合十分灵活，新的词语可以随意产生，缺乏一个完整的、公认的“词汇”体系。对拼音体系的外文来说，每一个词语怎样拼，是不是拼错了，有词典可查；但对汉语来说，除了政治经济用语基本上定型之外，生活用语和文学语言的字词组合简直千变万化，没有一个共同的标准，有些作家还要故作高深，故弄玄虚，正如鲁迅先生说的那样：“生造一些除了自己之外谁也不懂的形容词之类。”但就是鲁迅本人，也喜欢用些与众不同的词语：别人写“介绍”，他偏偏要写“绍介”，别人写“大致”，他偏偏要写“大抵”。因此，有人说中国只有“正字法”而缺少严谨的“正词法”，这话是有一定道理的。还拿 IBM 公司的语音输入系统为例，这套软件只能识别 35000 条“通用词语”，超过这个数字的许多词语电脑人都不认识，记下来的无非是些错别字。用它录入一篇短新闻还勉强能够对付，想用它来录入一部小说，困难就很大；想用它来写科技文章，哪怕没有数学公式或化学公式，就绝对不如键盘输入得心应手了。

第三，中国方言复杂，电脑语音录入，只能以普通话为标准(以后也许会增加广州话、上海话)，不可能懂得全国所有的方言。这就要求录入的文件，语言必须是规范的，也就是词汇和语音都是标准的，带有口音的人，就无法使用。

第四，用语音输入法口授一篇简单的通知书或信件之类还可以，想用口授的方法“说”出一篇小说来，恐怕不是大多数作家们所能胜任的。我是个作家，我能够写出一部百万字的长篇小说，但我绝“说”不出一篇只有一万字的小说来。因为小说是经过不断的推敲、修改“写”出来的。即便是说评书、讲故事，很可能也是事先准备了文字底稿，临场再即兴发挥罢了。

此外，用语音录入所产生的文件，据我所见到的，大都错别字连篇，而且通篇没有一个标点(语音录入能够识别标点，但几乎没有一个人习惯于说话中带标点符号)，说的人倒是痛快了，整理这种文件的人，可就叫苦连天。因此，可以这样说：语音录入，除了用作演讲记录的辅助工具之外，是永远不可能代替键盘录入被多数人所采用的。

手写录入，则是通过一种特制的电子笔把文字写在特制的书写板上，输入到电脑里，经过识别和内码转换，变成文字显示在屏幕上。但也有许多限制：第一，目前的手写识别，还只能做到大字识别，所写的字，至少要像核桃大小，小了就不认识；第二，字体必须工整而规范，潦草的或不规范的，系统就不认识。再说，正因为汉字书写缓慢而困难，所以才要电脑化，才要使用电脑来处理文字；配备了先进的电脑，还要用落后的笔来一个个书写，岂不是走了回头路了？因此，除了懒得学键位指法的老先生之外，对年轻人来说，这一功能缺乏先进性，发挥不出电脑的优越性，吸引力并不大。

第三种，称为“扫描输入，自动识别”，是用扫描仪先把文字扫描成一个图形文件，再用识别软件进行识别，转换成可以编辑修改的文本文件。这一方法，只要所写字体是清楚的、工整的，不要求写大字，写成小字就可以。当然，最有把握的是扫描印刷体。也就是说：举凡报刊书籍，都可以进行识别转换。这一功能，对于做文字资料工作的人来说，用处极大。特别是对作家、编辑、记者、教师等主要与文字打交道的人来说，可以用这个办法输入“电脑化”以前发表或出版的文章，然后进行修改。再版一部书或编一本个人集子，就用不着重新录入文字了。

为此，我们这套书不介绍语音和手写输入，而介绍扫描输入法，同时也简单介绍图片的扫描输入，想来对已经投身文字工作或打算从事文字工作的朋友一定有些用处。

第一章

图文扫描与字符识别

把图片用扫描仪扫描，变成一个图形文件，既可以编辑修改，也可以打印或以传真的形式发送，这属于图片屏幕编辑范畴。现代的新闻、出版、印刷行业，都要用到它。

用扫描仪扫描文字，不论是印刷的还是手写的，扫描的结果，是把一页文字变成一张图像，需要修改，也只能像图片一样，用图像编辑软件进行放大缩小或局部的修补。换句话说，通过扫描仪扫描所产生的文件，是图像文件，所有的文字是连成整体的，不是一个一个的，不能以文字编辑的方法进行替换或增删。

印刷品中的文字，通过扫描仪以图像的形式输入电脑，再通过识别转换，变成可修改的文本，这一过程，称为 OCR(是英文“光学字符识别”的缩写)。通过 OCR 处理产生的文件，是文本文件，所有字符都是“活”的，都可以用另一个字符替换或进行增删修改。

OCR 文字处理，属于办公自动化系统。

在西方，凡是使用拼音文字的国家，使用扫描仪把印刷品复制成可修改的文件，是早就已经实现了的事情。因为拉丁字母体系的拼音文字只有 26 个字母，使用斯拉夫字母或阿拉伯字母的文字，最多不过三四十个字母，加上大小写的区别和一些标点符号或特殊符号，也不过百把十个字符，让电脑识别和转换，都不是太难的事情。

汉字就不同了。汉字是在象形文字的基础上发展起来的，

数量众多，当代通用汉字，就在一万个上下。要请电脑认识这样多的汉字，即便不是不可能的事情，也是相当困难的事情。

自从汉字键盘输入法这个“瓶颈”问题相对完善地得到解决以后，清华大学紫光集团的专家们就开始攻“电脑识别汉字”这个难关。经过十几年的努力，不能说已经完全彻底地成功，但仅就清华紫光 OCR “通用汉字识别率达到 98~99%”这一成果，就已经很不简单了。

使用扫描技术进行图像处理和高速录入汉字，需要有硬件设备和软件技术两方面的配合。硬件方面需要一台高质量的扫描仪，软件方面需要一种高识别率的 OCR 系统。下面分别叙述。

一、硬件设备：扫描仪

扫描仪的种类很多，从外形与使用方式分，有手持式与平台式两大类；从分辨率分，有高分辨率与低分辨率两大类；从产地分，有进口与国产两大类；而国产扫描仪中，又分大陆产品和台湾产品两大类。

进口扫描仪，有 AGFA 系列、HP 系列、UMAX 系列、ANATECH 大幅面系列、PRIMAX 手持系列等等。国产扫描仪中，台湾是生产扫描仪的主要基地，厂家众多，产品丰富，在世界扫描仪市场上占有极主要的地位。大陆的扫描仪生产厂家是清华大学紫光集团，而且是大陆目前唯一的一家。所产的 UNISCAN 系列扫描仪，在国内外享有极高的声誉。

手持扫描仪价格低廉(800dpi 单色的，以前售价近千元，现在只卖四五百元)，但是扫幅仅有 10cm，只能用来扫描一些小件，扫描对象稍微大一些，就无法胜任。加上各种平台式扫描仪的价格近来不断降低(UNISCAN5A 和 4A 彩色扫描仪的售价不过 1500 多元)，手持式扫描仪大有被淘汰的趋向。因此本书只介绍平台式，不介绍手持式。

(一) UNISCAN 系列扫描仪的种类

UNISCAN 系列, 目前有“小旋风”系列的 5A、5B、5C 和 4x 系列的 4A、4B、4C、4D 等几种。

UNISCAN-5A 扫描仪, 是 1997 年 4 月上市的新产品, 光学分辨率为 $300 \times 600\text{dpi}$ (dpi 是 Dots per Inch 的缩写, 指每英寸的点数, 是扫描仪也是印刷制版和打印的分辨率单位); 最大分辨率为 $4800 \times 4800\text{dpi}$, 扫描面积为 A4 纸。它与 4-x 系列的最大不同之处, 是不再使用在扩展槽中加插卡的传统做法, 而改用 EPP (是英文 Enhance Parallel Port 的缩写, 意思是“增强并口”) 标准。也就是说, 扫描仪的电缆线不是插在插卡的接口上, 而是通过一只“软件狗”插在并口也就是打印口上 (这种“软件狗”上还有一个串联的接口, 可用来插打印机的电缆, 与打印系统并不冲突)。这样做有两方面的原因和好处: 第一是电脑内部扩展槽本来就不多, 一般最多只有七个, 除了必不可少的显示卡、多功能卡 (586 以上电脑不用) 之外, 有的人还要加声音卡、解压卡 (586 以上电脑可以不用), 这就占了四个扩展槽了。清华 OCR 正版软件还要装一块加密卡, 这就是五块卡了; 如果机器里还装有汉卡、语音校对卡等等, 扩展槽就不够用了。为了省出一个扩展槽来让别的系统使用, 扫描系统再也不能占用两个扩展槽了 (当初设计的时候就应该把两块卡合二为一的; 现在的新型并口“软件狗”已经兼有了接口卡与加密卡两种功能, 扩展槽上可以不插任何插卡了)。第二, 笔记本电脑里面没有扩展槽, 为了让笔记本式电脑也能接上扫描仪, 这就非改革接口方式不可。第三, 扫描仪使用插卡, 采用的是单向传输方式, 传输率最高只能达到每秒 150Kb, 改用 EPP 新型并口, 采用双向、半双工的数据传输, 最高传输率每秒可达 2Mb, 比传统的方式提高了十几倍。那可真是“王奶奶与马奶奶——相差的可不是一点两点”啦。而 UNISCAN-5A 的价格, 只比 4A 高 100 块钱。因此使用 586 机型的用户, UNISCAN-5A 是最佳的选择。当然, 如果您使用的是 486 机型, 那就只能望“仪”兴叹了。

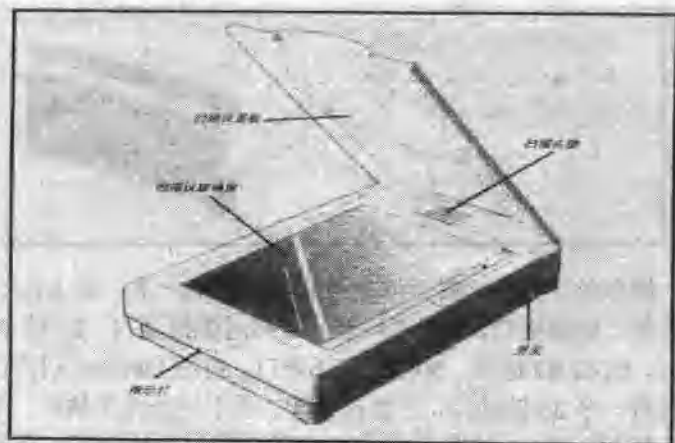
因为 UNISCAN-5A, 是不能使用于 486 以下机型的。

UNISCAN-5B 扫描仪的光学分辨率为 $400 \times 800\text{dpi}$; 5C 则为 $600 \times 1200\text{dpi}$, 扫描面积为 A4 加长。UNISCAN-4x 系列的光学分辨率, 4A 为 $300 \times 600\text{dpi}$, 4B 为 $400 \times 800\text{dpi}$, 4C 和 4D(专业级)为 $600 \times 1200\text{dpi}$; 扫描面积前两种为 A4, 后两种为 A4 加长。4-x 系列只要求 IBM PC/AT 386 以上机型(包括 586), 4M 以上内存和 10M 的硬盘空间。当然, 运行的速度比起 5 A 来, 可就差得远罗!

(二) UNISCAN 的安装

下面以 UNISCAN-4A 彩色扫描仪为例, 说明具体的安装方法。因为 4A 是最普及、最通用的机型, 使用的人也最多。一般说来, 只要会用 4A, 5A 也就不在话下了。

打开包装箱, 首先检查机件及附件是否完整齐备。A4 扫描仪的外形如下图所示。



图中的字太小, 看不清楚, 下面再说明一下: 顺时针方向, 从右上角开始, 分别是: 扫描头锁、开关、指示灯、扫描仪玻璃窗、扫描仪盖板。掀开机盖, 扫描头锁的箭头应该对准锁定