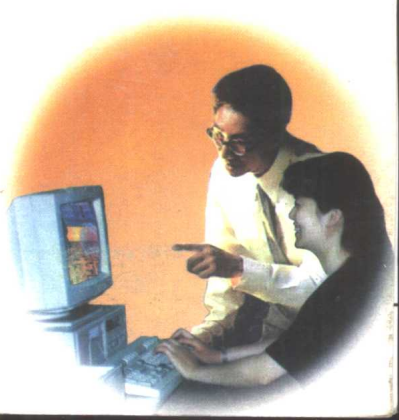
软件村

软件村——办公系列

扫描识别专家

OCR 7.0

《软件村》编写组



出版社

(京)新登字039号

京工商广临字98139

《软件村》丛书包括下列12个系列

◇办公系列	◇编程语言	◇操作系统
◇多媒体开发和工具	◇工业设计应用	◇实用小工具
◇数据库系列	◇图形图像工具	◇网络工具
◇系统检测与维护工具	◇压缩工具	◇游戏系列

软件村/办公系列

扫描识别专家OCR 7.0

《软件村》编写组编

策划编辑：张文虎　郎红旗

组　　织：Write Express

责任编辑：武志怡

封面设计：于　兵

*

化学工业出版社出版发行

（北京市朝阳区惠新里3号　邮政编码100029）

新华书店北京发行所经销

化学工业出版社印刷厂印刷

开本 787×1092毫米 1/32　印张 1　字数 23千字

1998年7月第1版　1998年7月第1次印刷

ISBN 7-5025-2166-6/TP・110

定价：3.00 元

《软件村》丛书目录

办公系列
电脑秘书Outlook
电子表格先行者Lotus 1-2-3
经典字处理WordPerfect
排版专家PageMaker
扫描识别专家OCR 7.0
数据管理Access 97
文字处理大师Word 97
演示能手PowerPoint
优秀电子表格Excel 97
中文字处理WPS 97
中文语音输入IBM ViaVoice
中文字表处理CCED

编程语言
编程第一站QBasic
可视编程Visual C++ 5.0
可视编程Delphi 3.0
可视编程Visual Basic 5.0
可视编程Visual J++ 1.1
媒体大师DirectX
网页的语言HTML 3.2

操作系统
经典操作系统DOS 6.22
DOS的门神Config.sys和Autoexec.bat
DOS中文平台UCDOS 7.0
微软贵族Windows NT
Windows中文平台UCWIN
Windows中文平台南极星和地球村
Windows中文平台四通利方
Windows中文平台中文之星2.5
风靡全球的Windows 95
世纪终极版Windows 98
装机必备

多媒体开发和工具
MP3豪放Winamp
"超级解霸"和XingMPEG Player
电影剪辑师Adobe Premiere
多媒体开发工具Authorware
多媒体开发工具方正奥思
腾图影视

工业设计应用
数学家MathCAD
电路设计工具Protel
电路设计工具Tango
工程设计师AutoCAD
运算工具包MATLAB

实用小工具
安装程序制作工具InstallShield
电脑保护神Norton Utilities
电脑工具箱PC Tools
高速复制工具HD-COPY
图标大师Microangelo
微软的拼音输入法
超级编辑器UltraEdit
整人专家FPE

数据库系列
数据库ORACLE
数据库FoxBASE
数据库FoxPro
数据库SQL Server

图形图像工具
灵巧图像处理工具Paint Shop Pro
三维动画3D Animation Pro
三维设计大师3DS MAX
图像处理大师PhotoShop 4.0
图形设计师Adobe Illustrator
图形设计师Corel Draw
反利照片伴侣Ulead Photo Assistant
图像浏览器ACDsee

网络工具
网景浏览器Netscape Communicator
网络探险家Internet Explorer 4.0
网上传输带CuteFTP
网上浏览第一家Mosaic
网上下载工具GetRight
网页制作宝PageMill
网页设计FrontPage 98
新型的离线浏览器WebZip
主页编辑器Claris Home Page

系统检测与维护工具
WIN95的减肥茶CleanSweep
磁盘碎片整理能手Diskeeper
电脑保安KV300和KILL95
杀毒工具行大98和TBAV
杀毒能手AV95
系统测试工具WinBench

压缩工具
大众压缩软件Arj-WinArj
压缩多面手WinRAR
神奇压缩工具ZipMagic
压缩宠物WinZip和PKZIP-PKUNZIP

游戏系列
FIFA足球经理
阿猫阿狗
暗黑破坏神DIABLO
地雷战
帝国时代
古墓丽影II
红色警报Red Alert
"街霸Zero"和"VR战士"
金庸群侠传
三国志孔明传
篮坛霸主NBA 98
魔法军团
魔法门之英雄无敌II
三国游戏纵横谈
世界杯之路FIFA 98
天龙八部
网络游戏MUD
象棋大师
星际争霸Star Craft
Windows桌面游戏

12个系列 ● 100种 ● 每本3.00元

一个集电脑字典、全屏汉化、机器翻译为一体的集成软件系统

朗道电脑字典 翻译系统 VER 5.0多媒体光盘

● **专业化的电脑字典** 包括通用、电脑、医学、化学化工、电子、经贸、机械、建筑、法律在内的十余个英汉、汉英双向专业字库，词汇量超过百余万条，利用自建字库功能，用户可无限扩充字库。

● **地道的美语真人发声** 多达6万余条的英文词汇真人发声，特聘美国电台男播音员录制，标准美语，音质清晰。

● **全面、方便的字典查询功能** 自带中文平台，DOS / Windows 3.X / Windows 95 / Windows 98 / Internet全兼容，可选择鼠标即指即译，也可选择鼠标捕获单词翻译，还可键盘输入单词翻译，同时还可选择英文词汇同步发音，只需按一下鼠标右键即可将中英文翻译结果回送到当前编辑器中，方便地实现译文的校核。

● **基于人工智能的全屏汉化** 只需鼠标一点即可瞬间将屏幕上的英文全部翻译成中文，包括任意多级菜单、对话框、提示信息、帮助及编辑正文。先进的人工智能技术和详尽的语法分析使翻译结果的可读性大大提高。

● **文章翻译系统** 可将您选定的整篇英文文章或英文资料翻译成中文文章存盘，也可有选择地形成中英文对照的文章。同时，利用朗道字典的查询功能和翻译回送功能，可比较多义词的细微差别，对译文进行精细加工，从而提高翻译的质量。

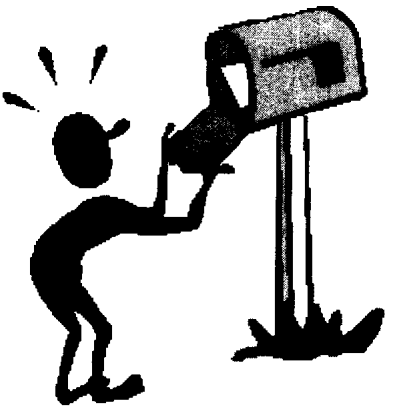
上海朗道电脑科技发展有限公司
上海浦东德平路12弄10号大楼304室
邮编码：200135
电话/传真：(021)58606142、58214389
E-MAIL：langdao@public.shanghai.cngb.com

全国各地各大软件经销商处有售

在办公自动化日益普及的今天，把文档资料输入计算机已成为企业日常事务中不可缺少的一部分。在过去，这一工作通常由那些心灵手巧的打字小姐（或先生）来完成。毫无疑问，这是一项简单但却十分枯燥而繁重的工作。更重要的是，它占据了企业运转中大量宝贵的时间（商业社会里时间可就是金钱喔！）。而且，万一碰上紧急情况，比如说某重要的国际商务会议突然提前召开，老板要某打字小姐在半天

内把厚厚的一摞资料输入笔记本电脑里，以让他在上飞机时带走。这时，有一点绝对可以肯定，即使这位小姐长有如孙大圣般的三头六臂，也还会不知足的埋怨爹娘为啥不把自己生成一个千手观音。扫描仪的出现缓解了这一矛盾。扫描仪可以通过光电效应把纸张上的印刷符号（文字和图形）变成图像存储在计算机里，从而可以极大的提高资料的输入速度。（打字小姐和先生可以长出一口气了）。但是，扫描进计算机里的只是一幅幅往往带有一点儿瑕疵的图片。而我们所需要的却常常是清晰、整洁的文档。更令人恼火的是，我们无法像处理文本一样对图像中的文字符号进行方便的编辑和修改。您也许会问："难道我们只能无可奈何的接受这个残酷的现实吗？""究竟有没有一种能方便的把字符从图像文件中挑选出来变为文本文件的工具呢？"。人，想得到就做得到！清华文通 TH-OCR 正是干这活儿的行家！

下面就让我们开始见识见识这位文字识别专家吧！



1 安装清华文通 TH-OCR

安装清华文通OCR是一件非常简单而轻松的工作。首先，你要将装有安装程序 Setup.exe 的磁盘或光盘放进相应的驱动器。如果您是在 MS-DOS 环境下进行操作，您需要：

(1) 进入 Setup.exe 所在的目录；

(2) 在该目录下键入 Setup，然后回车。

当然，您也完全可以键入完整的路径名直接运行 Setup.exe 命令。其实在软件的使用处处追求界面友好的今天，绝大多数的计算机上都已安装了 Windows95 系统。因此，下面我们将主要介绍 Windows 95 环境下清华文通 OCR 的安装情况。

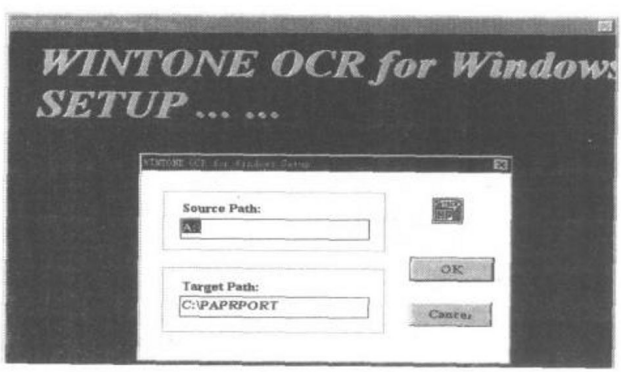


图 1-1 TH-OCR 的安装界面

（1）您需要双击打开 TH-OCR 安装程序所在的文件夹。

（2）然后双击 Setup.exe ，进入清华文通 OCR 的安装画面，如图 1-1 所示。

（3）在 Source Path(源路径)框中显示的是清华文通 TH-OCR for Windows 安装文件所在的路径，而在 Target Path（目标路径）框中显示的是安装这一软件的路径，也就是说，我们在硬盘上放置清华文通 OCR 的地方。当然，如果实际路径与上述路径不符，您完全可以对它们作出修改。确定路径以后，单击 OK 按钮，进入下一个步骤。

（4）这时，屏幕上会出现文通 OCR 快速安装的情况，这个状态会持续几十秒钟。在这过程中，如果您想放弃安装，则可以用鼠标单击 Cancel 按钮，退出安装。

（5）文件拷贝完毕后，屏幕上会出现一个窗口告诉您文通 OCR 已安装成功。按下“确定”按钮，屏幕上出现注册菜单，要您输入“姓名”和“单位名称”，输完后单击 OK 按钮（您也可以不进行注册，直接按下 OK 按钮）进入下一步。

（6）执行完上述操作后，屏幕上会出现一个对话框，如图 1-2 所示，问您是否愿意为清华文通 OCR 建立一个程序组。

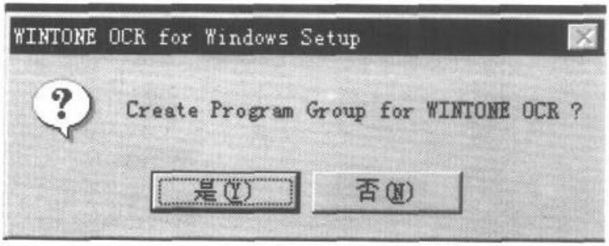


图 1-2 是否为 WINTONE OCR 建立程序组

如果单击  按钮，安装程序就会在 Windows 95 的开始菜单  中为 TH-OCR 争取到一席之地。

到此为止，清华文通 TH-OCR 算是真正在您的计算机上落户了。图 1-3 所示即为它在开始菜单中的快捷方式。

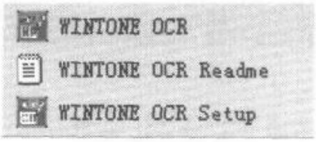
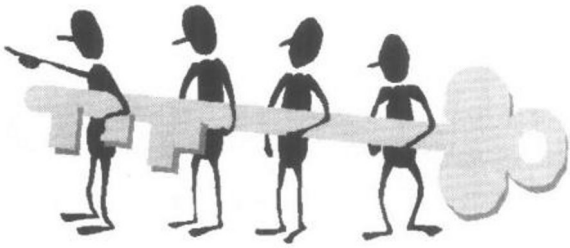


图 1-3　TH-OCR 的快捷菜单



2 清华文通 TH-OCR 的发展及其功能特点

2.1 清华文通的发展历程

OCR 是英文 Optical Character Recognizer 的首字母缩写，意思是光学符号识别器。在国内市场上品牌众多的 OCR 中，清华文通算得上是实力雄厚的佼佼者。从 1989 年第一版 TWReader 呱呱坠地到 1996 年推出清华 Wintone OCR for Windows，文通 OCR 的技术得到了长足的发展。1998 年 2 月，清华文通 TH-OCR MF 7.0 版又傲然出世，标志着文通 OCR 的技术又上了一个新的台阶。

2.2 清华文通的卓越性能

作为一个成功的中英文 OCR 系统，清华文通 TH-OCR 与国内外同类中英文 OCR 系统相比较，具有突出的特点：

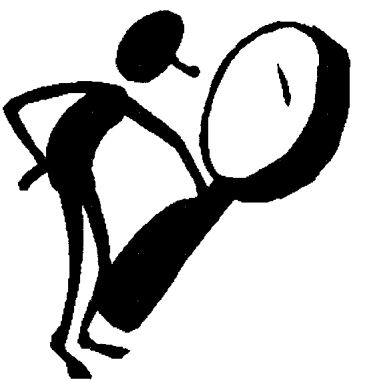
- 首创“汉英双语混排”同时识别功能，识别率最高，居国际领先水平。
- 首创简繁体汉字和英文的“多种字体混排”同时识别。
- 独家支持 Windows 环境下的“多种汉字内码”，适合全球各地区使用。
- 独家支持将识别结果“自动送入其他的应用程序和剪贴

板”，十分方便。

- 提供核心模块的开发接口，允许您使用识别技术开发自己的应用系统。
- 强大的批处理功能，可以同时处理多页文档资料，录入效率自然高人一筹。

而最新推出的 TH-OCR MF 7.0 更首创“非特定人脱机手写体汉字识别功能”，规范手写体的识别率可达 88 %~95 %以上。因此，那些厌恶敲击键盘的朋友也不用再发愁了！ TH-OCR MF 7.0 拓展了 OCR 系统的语言识别范围，首创日文和日汉混排功能，解决了需要进行日文资料输入的用户的燃眉之急。

在拥有了强大实力的前提下，文通 TH-OCR 的使用却十分简便。任何人都可以在经过短暂的学习后熟练的进行使用。文通 TH-OCR 的用户界面简洁、友好。下面，就让我们来一起学习文通 TH-OCR 的使用方法。



3 清华文通 TH-OCR 使用快速入门

3.1 清华文通友好的用户界面

清华文通 TH-OCR 的工作界面分为识别部分和编辑部分两个层次。其中识别部分的界面如图 3-1 所示。

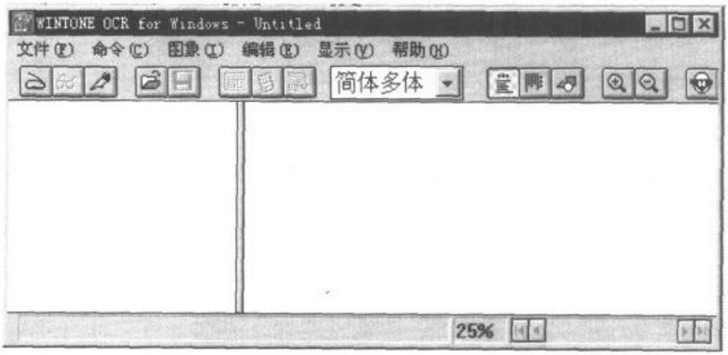


图 3-1 TH-OCR 的识别窗口

识别窗口的最上方是标题栏。标题栏的末尾是打开的文件名。文件名随打开文件的不同而发生变化，但默认的缺省值是“Untitled”。窗口中在下方紧挨着标题栏的是识别菜单条。识别菜单中的选项包括了文字识别所需的所有功能。在菜单中有一些变为灰白的选项，表示这是“编辑”菜单的

功能，只能在编辑窗口中使用。如图 3 -2 所示。

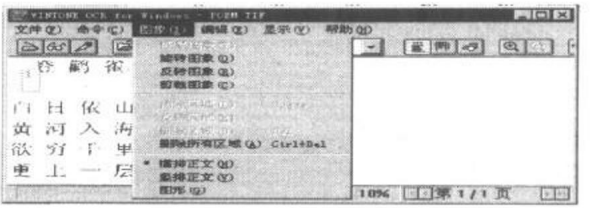


图 3-2 识别菜单选项示例

在识别菜单的下方是识别工具条。如图 3-3 所示。

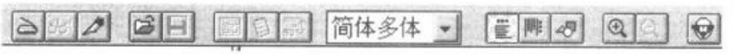


图 3-3 识别工具条

在识别工具条上有许多按钮。它们是识别菜单中各选项的快捷使用方式。换句话说，那些需要您先打开菜单，再将

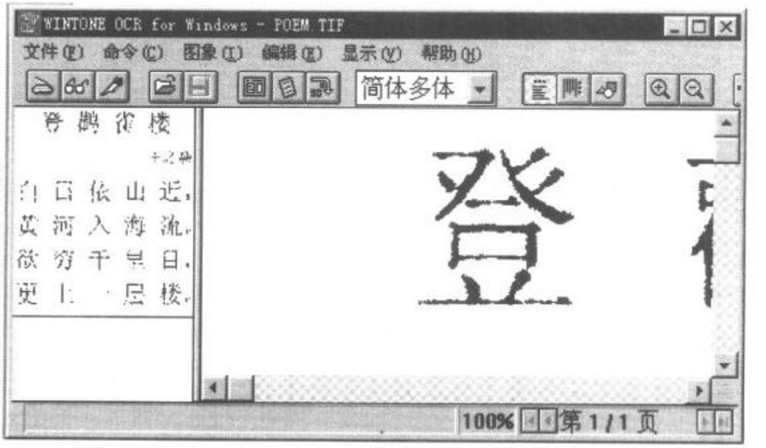


图 3-4 识别窗口示例

光标移到选项上，然后单击该选项才能实现的功能，您只需单击快捷按钮便可实现，何乐而不为呢？

识别窗口的主体部分被划成了左右两部分，如图 3-4 所示。

左边的窗口较小，用来观察所打开文件的全貌；右边的窗口较大用来观察文件的局部并对文件进行识别前的调整。

> 提示：您可以用拖曳两窗口之间分隔条的方法来调整两窗口之间的比例大小。

在窗体的底部是一个状态条，它显示了当前图像的缩放比例和当前所打开的文件的页数信息。例如，图 3-4 的状态条显示的就是当前图像放大比例为 1，目前程序共打开如图所示的一页图像。

编辑窗口与识别窗口十分相似，它也是由标题栏、菜单条、工具条、操作窗口和状态栏组成。如图 3-5 所示。

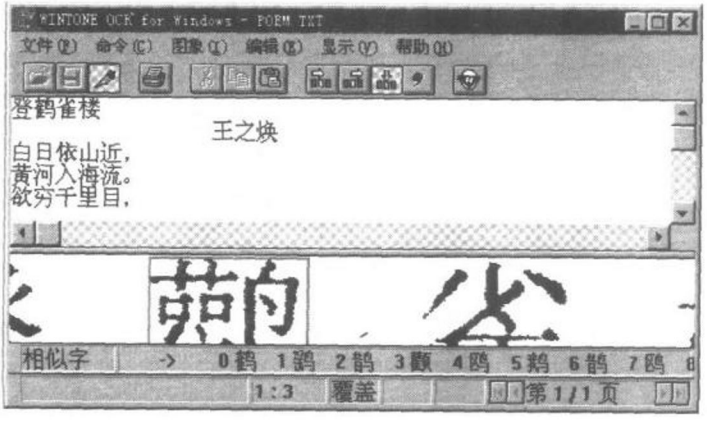


图 3-5 编辑窗口示例

编辑窗口与识别窗口的菜单条外表上一模一样，但菜单内部的具体选项却有很多不同。原先在识别窗口下不能使用的选项现在可以使用了，而先前保持激活的一些选项现在却变成了灰白色。与之相应，编辑工具条上那些作为菜单选项快捷途径的按钮也相应的发生了变化。在工具条下方，是一个由分割条分开的两个窗口。上窗口（或左窗口）用来展示和编辑经过识别后的文本，下窗口（或右窗口）用来显示原图像文件。编辑菜单底部的状态栏多出一行待选的字符，它让用户从中选择输入正确的文字或常用符号。

提示：细心的读者也许会注意到，在图 3-4 和图 3-5 中，标题栏里文件名的扩展名发生了变化，由*.TIF 变为了*.TXT 。这并不奇怪。在识别窗口里打开的是图像文件，理应以表示图像文件的扩展名 TIF 结尾；而在编辑窗口里打开的却是已经过识别处理的文本文件，自然应该用表示文本的扩展名 TXT 结尾了。

编辑窗体中，文本窗口和图像窗口究竟是上下分布还是左右分布依赖于在图像菜单中设定的识别区域是 竖排正文(V) 还是 横排正文(H) 或者设定为 图形(G) 。

3.2 清华文通强大的功能

清华文通 TH-OCR 所具有的强大功能体现在它对图像文件灵活、高效的处理方式和对识别出的文本文件周密、细致的编辑方法上。当然，那极高的识别率就更不用说了。下面，我们就把所有菜单选项的功能和工具条上相应的快捷按

钮一一列举出来。

3.2.1 识别窗口

识别窗口有如下选项：

(1) “文件”菜单

“文件”菜单如图 3-6 所示。

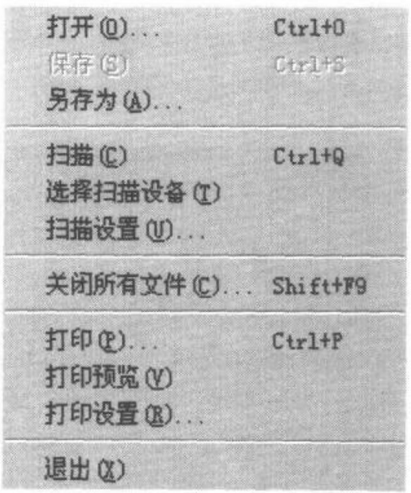


图 3-6 识别窗口的“文件”菜单

- 打开 从磁盘中打开并读入一个已存在的图像文件。
- 保存 将版面倾斜校正及版面分析的结果存盘。
- 另存为 将文件用新的名字存于新的文件夹里。
- 扫描 开始进行扫描。
- 选择输入设备 选择扫描设备。
- 扫描设置 设置进行扫描时所采用的界面、扫描仪的分辨率、扫描的页面长度、所用亮度等等。

- 关闭所有文件　关闭所有打开的图像文件。
- 打印　打印当前的图像文件。
- 打印预览　在开始打印之前预先浏览将被打印的文件。
- 打印设置　设置打印机、打印纸张等参数。
- 退出　退出清华文通 TH-OCR 系统。

（2）“命令”菜单

“命令”菜单如图 3-7 所示。

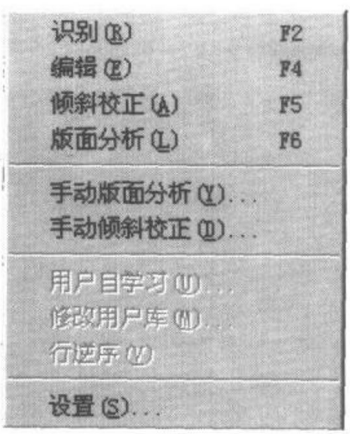


图 3-7　识别窗口的“命令”菜单

- 识别　开始根据设置的参数（如字体）等进行识别。
- 编辑　进入编辑窗口，对识别产生的文本进行分析。
- 倾斜校正　开始由系统自动对版面进行倾斜校正，片刻后显示窗中将显示校正后的图像。
- 版面分析　开始由系统自动对版面进行分析，划分出若干个方块区域，并赋给它们不同的属性。
- 手动版面分析　由用户自己动手进行版面分析。当用户