

TH-OCR®

TH-OCR

联想专业版2.0 使用手册

目 录

序言

第一章 TH-OCR 联想专业增强版的安装	(1)
1.1 系统的运行环境	(1)
1.2 系统软件的安装	(1)
第二章 操作流程	(4)
2.1 系统设置	(6)
2.2 获取所要识别文件的图像	(6)
2.3 进行图像版面处理	(6)
2.4 进行识别及修改	(6)
2.5 识别结果输出	(7)
第三章 功能详解	(8)
3.1 文件	(8)
3.2 图像	(18)
3.3 命令	(24)
3.4 编辑	(34)
3.5 显示	(38)
3.6 帮助	(40)
3.7 退出系统	(40)

序 言

“TH-OCR®清华文通多体中英文混排印刷文本识别系统”是由清华大学电子工程系研制开发,北京市清华文通信息技术有限公司总经销的高性能多体中英文混排印刷文本识别系统。该系统在印刷汉字及英文文本的输入工作中完全代替了人工键入。它速度快、准确率高、实用性强,各项指标在国内外均处于领先地位。目前已被国家机关、企事业单位及个人广泛应用于办公、印刷、出版、新闻、信息、金融、军事等各个领域,市场占有率高达65%;产生了巨大的社会效益和经济效益。

“清华TH-OCR®联想专业增强版综合集成文字识别系统”是“TH-OCR®清华文通多体中英文混排印刷文本识别系统”在Windows环境下的最新版本。它保留了原有版本简、繁体多体汉字识别、中英文混排同时识别、横竖版面混排及复杂版面分析、理解与文字识别、表格自动识别、批量识别及将识别结果自动送入其它应用程序和剪贴板等功能外,新增加了手写体识别、灰度彩色图象识别、识别结果版面恢复和保持等功能,因而应用面更广,实用性更强,同时识别的速度及正确率也有一定提高,是清华TH-OCR®文字识别领域的新的里程碑。

有关清华TH-OCR®汉字识别技术的特点、发展历程和其它详细内容,请参阅清华大学出版社出版发行的《清华TH-OCR®技术应用与开发》一书。

声明:

TH-OCR®的技术支持及软件维修均由北京清华文通信息技术有限公司负责。TH-OCR®是我公司的注册商标。请认准商标谨防假冒。

北京清华文通信息技术有限公司
2001年04月

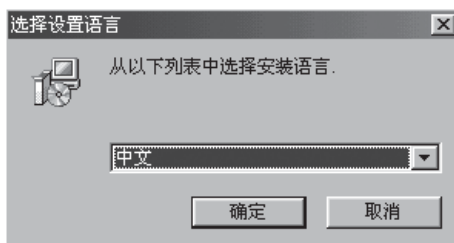
第一章 TH-OCR 联想专业增强版的安装

1.1 系统的运行环境

- a) 586 以上档次微机 (推荐使用奔腾 MMX166 以上微机)。
- b) 32MB 以上内存 (推荐使用 64MB 以上内存)。
- c) 1GMB 以上硬盘。
- d) 支持 TWAIN 标准的扫描仪。
- e) 中文 Windows 95 以上版本或使用带有中文之星、四通利方等汉字环境的西文 Windows95 以上版本 (包括 Windows95、Windows98、WindowsNT4.0、Windows2000 及 WindowsMe)。

1.2 系统软件的安装

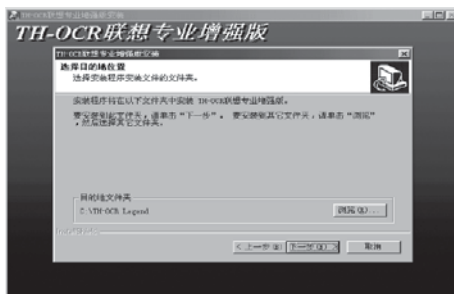
装入 OCR 光盘后, 系统自动启动安装程序, 并出现以下画面:



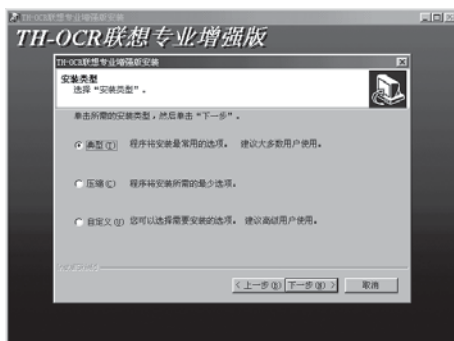
选择确定后, 则出现以下界面:



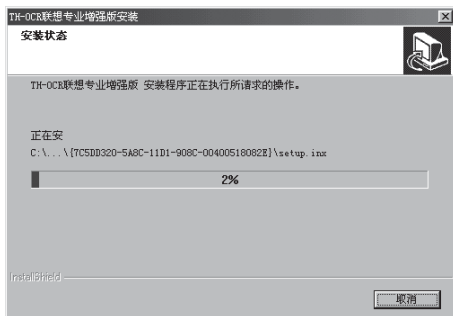
等完成后，选择确定，进入以下画面：



选择下一步，进入以下画面：



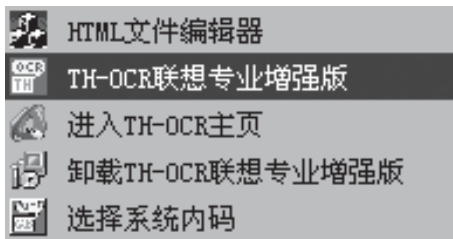
可以根据需要选择，这里以选择典型安装为例，选择下一步，进入以下画面：



软件安装开始，完成后，选择确定，进入下一步：



选择所需要识别的字符集，按OK后，软件就安装完成了，在程序中形成本软件的图标，如下图所示：



第二章 操作流程

TH-OCR 联想专业增强版，操作流程分为五步：

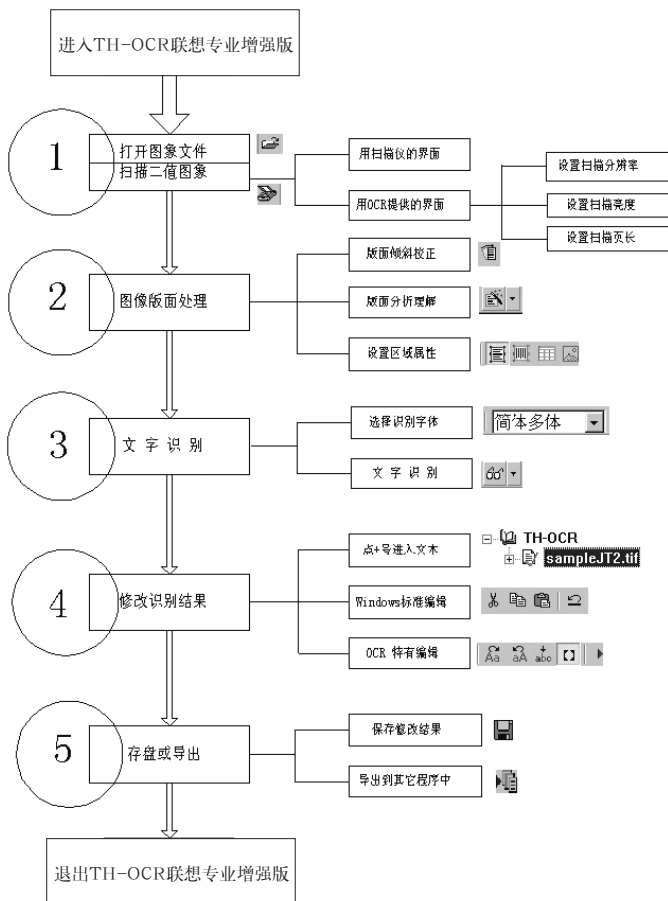


图 1 是系统主界面的构成。

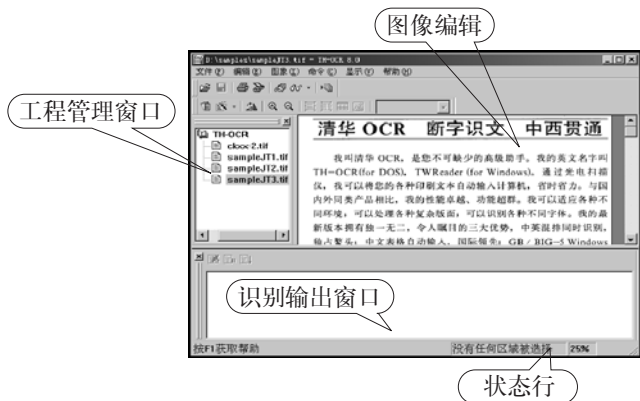


图 1-1 图像编辑界面

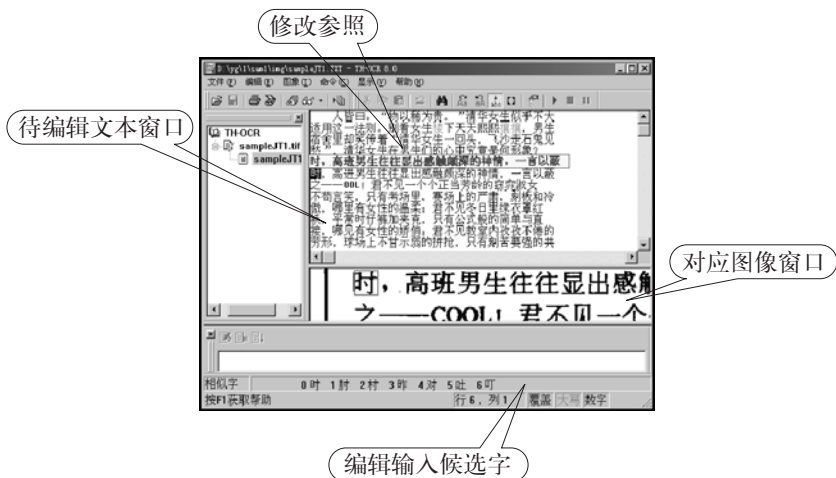


图 1-2 文本编辑界面

2.1 系统设置（详见 3.3）

使用系统前应根据应用环境及需求设定系统参数。打开【命令】中的【设置】项，可对系统、识别、后编改、语音校稿及其它参数进行设定。设定结果系统会自动记录并保持，不必每次操作都重新设置。

2.2 获取所要识别文件的图像（详见 3.1）

对于磁盘上原有保存好的图像文件，可以直接用【打开】命令或工具条上的打开；如果要获取新的图像文件，则应选择【扫描】命令或工具条上的，对所要识别的文件进行扫描。

注意：

- ★ 扫描仪的操作应遵照随同扫描仪发送的操作手册的指导；
- ★ 对普通书本的印刷质量，字号在5号以上的印刷材料，可适当选择其扫描分辨率为 300，扫描亮度为 0。
- ★ 对已有的图像文件，要注意其图像存储格式是否符合TH-OCR 联想专业增强版系统的要求（TIF 格式、BMP 格式、JPG 格式或 PCX 格式）。

2.3 进行图像版面处理（详见 3.2、3.3）

对扫描所得图像文件根据需要进行处理（旋转、反转、剪裁、倾斜校正、版面分析……等等），为识别做好必要的准备。注意选择要进行文字识别的区域，如果不选择，则认为是对整篇图像进行识别。在对整篇图像进行识别时，识别区域中不能包含有插图。

2.4 进行识别及修改（详见 3.3）

选择识别命令完成文字的识别。完成识别后，点击工程管理窗口中相应的 TXT 文本文件进入编辑修改即校对状态。

2.5 识别结果输出（详见 3.1）

识别结果经修改编辑后，可根据需要通过【导出】命令输出到其它应用程序中或直接存盘。

第三章 功能详解

TH-OCR 联想专业增强版系统的全部功能体现于主菜单的【文件[F]】、【编辑[E]】、【图像[I]】、【命令[C]】及【显示[V]】等命令中，每一个命令均带有一个子菜单。为了方便快捷的操作，还设置有对应的工具条（图2），常用的命令均可直接从工具条中选定。本章将详细介绍各子菜单中命令的功能及使用方法。

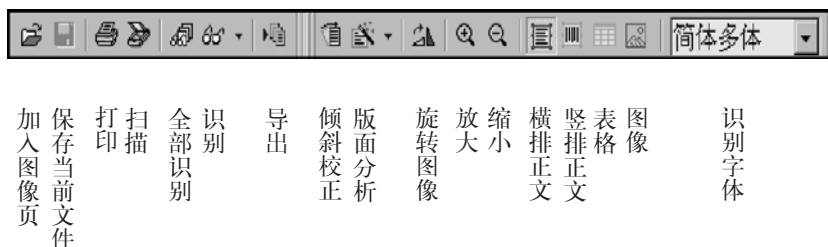


图2 图像编辑环境下的工具条功能

3.1 文件 [F]

【文件[F]】菜单中包括有14个命令，分述如下。如（图3）



图 3

3.1.1. 文件的打开与保存

在 TH-OCR 联想专业增强版系统中，图像文件的打开与保存等操作，采用 Windows 的标准风格。

a) 打开 [O]

在【文件[F]】菜单中选【打开[O]】，或用鼠标选工具条上的屏幕上显示【打开】对话框，如（图 4）所示：

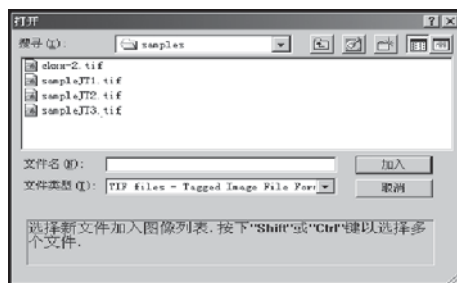


图 4

注意：当需要打开多页时请按住CTRL键后选中所有需要打开的图像页，图像文件则被打开。

3.1.2. 保存[S]、另存为[A]

图像可以用原名保存，也可以换名另存。用原文件名保存时，在【文件[F]】菜单中，选【保存[S]】项或在工具条中选即可；换名另存时，选【另存为[A]】项，屏幕显示出【换名存盘】对话框（图5）

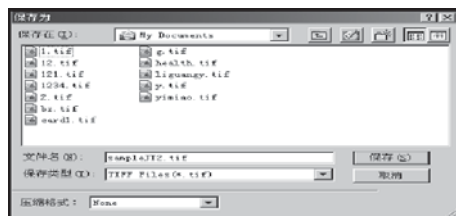


图5

3.1.3. 导出[E]

识别、校对完成后，对所获得的文本保存。（如图6）

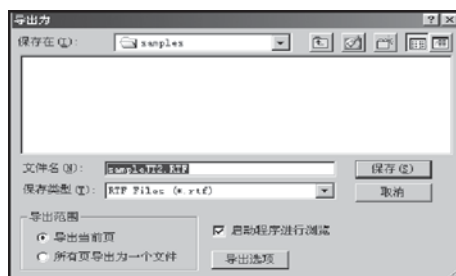


图6

- a) 选择保存路径，键入文件名。
- b) 【导出范围】：选择【所有页导出为一个文件】时，系统把所有识

别结果合并并导出。

c) 【启动程序进行浏览】被选中，在导出时系统自动启动与导出文件格式对应的浏览程序。

d) 点击【导出选项】，不同的文件类型有不同的选择。如图分别是RTF（图7）、HTML（图8）、TXT（图9）文本格式的导出选项。用户可根据需要自行选定。



图 7

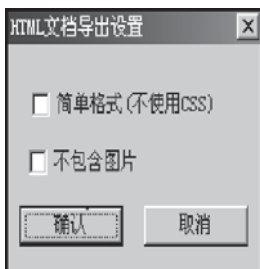


图 8

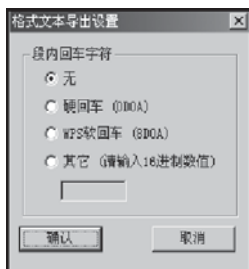


图 9

3.1.4. 选择扫描设备 [T]、扫描设置 [U]、扫描 [C]

扫描是获得待识别图像的手段。扫描前必须做好选择扫描设备、扫描设置的工作。

A. 选定扫描仪 [T]

当选定【选择扫描设备 [T]】项时，若计算机系统中已安装了扫描仪及其驱动程序，则出现 TWAIN 提供的【Select Source】对话框，选择所安装的扫描仪，然后按【Select】键确定。若尚未安装扫描仪及其驱动程序，则此选项无效。这时应按扫描仪安装要求进行扫描仪及其驱动程序的安装。扫描仪一旦选定，以后就不用再选择。

B. 扫描设置 [U]

在【文件 [F]】菜单中选定扫描设置 [U]】项。屏幕上出现【扫描设置】对话框。对话框中提供两种扫描界面供用户选择：

A) 使用 TWAIN 扫描界面

这是使用扫描仪自己的界面扫描图像，详细操作可参考扫描仪的操作

说明。此时对话框下部的选项无效。

B) 直接终扫

这是使用 TH-OCR 联想专业增强版自己的界面扫描图像。此时需要同时设定对话框下部所示的扫描参数：

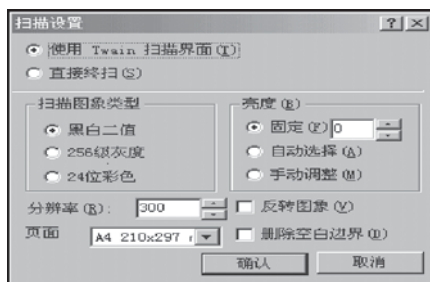


图 10

a 亮度

亮度参数是扫描仪的最重要的参数之一，也是影响 OCR 系统文字识别效果的极为重要的因素。在 TH-OCR 联想专业增强版系统中，提供了亮度确定的三种方式：如果选【固定[F]】，可在 中直接输入亮度数值或用鼠标按其右边的箭头，以获得所需的亮度数值。系统中亮度数值的调节范围是 -128 ~ 128。

如果选【自动[A]】，则亮度参数完全由 TH-OCR 联想专业增强版系统在扫描过程中自动确定。如果选【手动调整[M]】，则亮度参数由用户在扫描过程中根据具体情况调整确定。调整的方法详见本节3) 扫描部分。有一定经验的操作者，使用【手动调整[M]】可以获得比较理想的扫描识别效果。

注意：

亮度参数的确定决定了扫描图像的明暗程度，亮度太高，文字笔划断裂而残缺不全；亮度太低，文字笔划相互粘连而黑成一团。两种情况都会影响识别结果。因此，认真细致地选择好亮度是获得较高的识别率的基本保证。

有关亮度选择的几点建议:

(1)扫描较浅的文件或底色白、纸张发亮以及文字笔划细的文件时,亮度参数应选低些,即加黑、加暗;

(2)书刊、杂志的底色较白,宋体字、仿宋体字的笔划粗细适中,在识别这类文件时,可选用中等亮度。

(3)识别较小的文字时,亮度参数应高些,以不出现太多的断笔为限。

如图 11、图 12、图 13 分别为扫描的亮度太高、太低以及亮度相对适中所得到的三种图像的示意图。

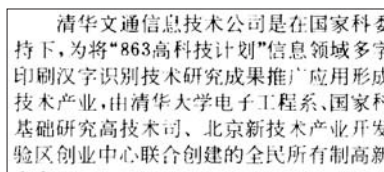


图 11

图像的扫描亮度太高,
文字笔划多处断裂。

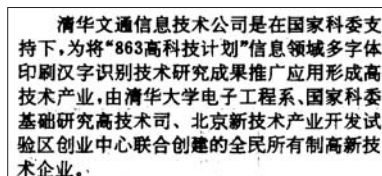


图 12

图像的扫描亮度太低,
文字笔划多处粘连。

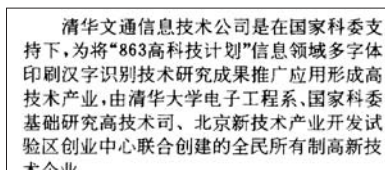


图 13

图像的扫描亮度相对适中,
文字笔划比较清楚。

b 分辨率

分辨率是扫描仪也是 TH-OCR 联想专业增强版系统的另一个重要参数。它决定了扫描图像的清晰程度,也决定了所能扫描的图像细节。文本

识别正确率的高低与扫描分辨率的设定有很大关系。扫描仪分辨率的单位是 DPI (Dots Per Inch, 每英寸点数)。可以在  中直接输入分辨率数值或用鼠标点按其右边的箭头, 改变分辨率数值。在 TH-OCR 联想专业增强版系统中, 分辨率选择的有效范围是 100 ~ 600 DPI。常用材料的最小字号是小 5 号 (9 磅) 或 5 号字 (10.5 磅), 选择 400DPI 或 300DPI 的分辨率, 可以适合大多数情况。分辨率的精细设定请参考下表。

分辨率选择对照表

文字大小□	准确分辨率(DPI)□	推荐使用的分辨率(DPI)
1号 (26磅)	150□	200□
2号 (22磅)	180*	
3号 (16磅)	200	
4号 (14磅)	240	300□
小4号 (12磅)	280	
5号 (10.5磅)	300*	
小5号 (9磅)	350□	400□
6号 (7.5磅)	400	
7号 (5.5磅)	500	600□
8号 (5磅)	600*	

选择 180DPI 时, 扫描仪扫描得到的图像与点阵打印机的打印结果为 1:1。选择 300DPI 或 600DPI 分辨率时, 扫描仪扫描得到的图像与激光打印机的打印结果为 1:1。

c 页长

平板扫描仪的扫描幅面一般是 A4 或长度为 14 英寸。有许多时候,扫描的材料可能幅面很小,为了节约扫描时间,提高工作效率,可以预先设定这一参数。在 TH-OCR 联想专业增强版系统中页长有效值是 6 英寸、9 英寸和 12 英寸。

d 反转图像

用于 TH-OCR 联想专业增强版系统的图像,有时是黑白二值图像(Black and White Drawing)。但是,扫描得到的黑白二值图像,因扫描仪型号的不同而各有差异,即有的图像可能是白底黑字,而有的则可能是黑底白字。在 TH-OCR 联想专业增强版系统中,能识别的图像应是白底黑字的图像。如果扫描后,在系统中看到黑底白字的图像,请将**反转图像**设置为有效,即 .

e 删除空白边界

绝大多数文件材料的边界都有很多空白,而没有任何内容。为了便于识别处理,提高系统的处理效率,可以将这项设置为有效,扫描时,系统会自动滤除图像边缘的空白。

在实际应用中,如何设置上述各参数,会因操作环境、要扫描的材料以及具体需要等多方面因素的不同而不同。只要按照以上说明,认真操作,自然会得心应手。

C. 扫描 [C]

完成各个扫描参数的设置后,在【文件[F]】菜单中,选定【扫描[C]】项或在工具条中按  键,开始扫描。

如果已在【扫描设置】中选择了【使用 TWAIN 扫描界面】,则要进行两次扫描。第一次预扫目的是测定被扫描文件的亮度,用户在预扫结果中对扫描亮度、分辨率及扫描范围进行选择调整后再进行终扫;

如果选择的是【直接终扫[S]】，则系统只进行一次扫描。假如用户在【直接终扫[S]】的亮度选择中已选【固定[F]】或【自动[A]】，扫描后将直接显示扫描获得的图像；假如选的是【手动调整[M]】，则扫描结束后，屏幕显示【选择亮度】对话框。

操作者可以一边调【选择亮度】对话框中的亮度调节图 14 条，一边观察窗口中图像的明暗变化和扫描质量，直到对图像满意；也可以选择对话框左下方的【自动[A]】键并观察图像效果。当得到满意的图像时，用鼠标按【确认】键，此时系统将退出【选择亮度】对话框并重写图像。

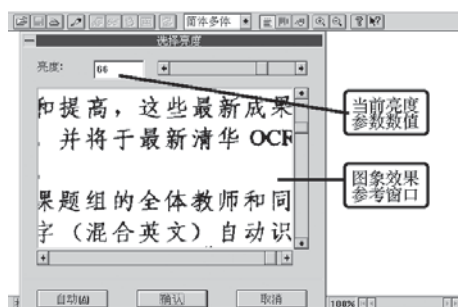


图 14

3.1.5. 打印 [P]、打印预览 [V]、打印设置 [R]

TH-OCR 联想专业增强版系统可以在图像环境或文本编辑环境中将扫描得到的图像或识别结果文本直接打印输出。有关打印输出，共有三个命令：【打印 [P]】【打印预览 [V]】及【打印设置 [R]】。

a) 打印设置 [R]

选定【打印设置[R]】项后，屏幕显示【打印设置】对话框，在选择打印机及设定打印参数后，按【确定】键退出。



图 15

b) 打印预览 [V]

选【打印预览 [V]】项后，屏幕显示模拟的打印图像。按 Alt-F4 键或选【关闭】键可以退出返回图像环境。



图 16

c) 打印 [P]

选定【打印 [P]】项后，屏幕显示【打印】对话框，供用户选择打印页码范围、质量以及打印份数等。

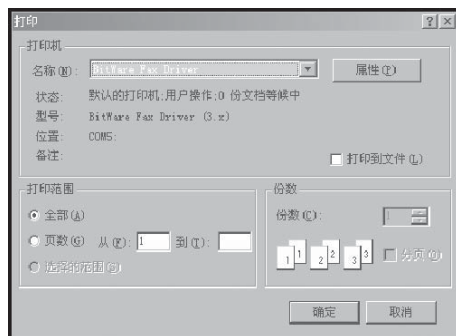


图 17

3.2 图像 [I]

在 TH-OCR 联想专业增强版 系统中，利用【图像 [I]】子菜单的有关命令可以对待识别的图像进行各种处理以便有效的识别。【图像 [I]】功能菜单如图 18 所示。



图 18

3.2.1. 反转图像 [I]

在【图像 [I]】菜单中，选【反转图像 [R]】项，可以实现图像的黑白反相。

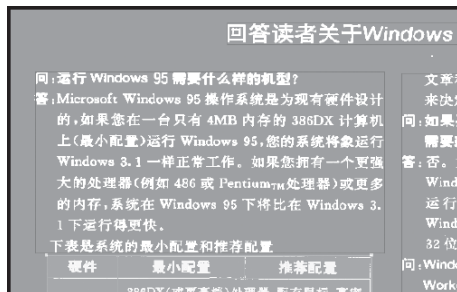


图 19

如果扫描得到的图像如图，为黑底白字（即文字是白的，背景是黑的），就需要进行黑白反相处理，才能成为可供识别的白底黑字图像。

3.2.2. 旋转图像 [O]

在【图像 [I]】菜单中，选定【旋转图像 [O]】或在工具条中按图像旋转键  一次图像将顺时针旋转 90°，如果连续多次选用该项，可以实现图像的 180°、270° 乃至 360° 旋转。图 20 为正确图像顺时针旋转 90 度后的结果



图 20

图像旋转功能的主要用途: 当所要扫描识别输入的材料宽度超过了扫描仪的宽度 (通常为 8.5 英寸) 时, 可以将材料旋转 90° (后), 放入扫描仪进行扫描。这样扫描得到的图像需要经过旋转, 恢复原样才可以进行识别处理。

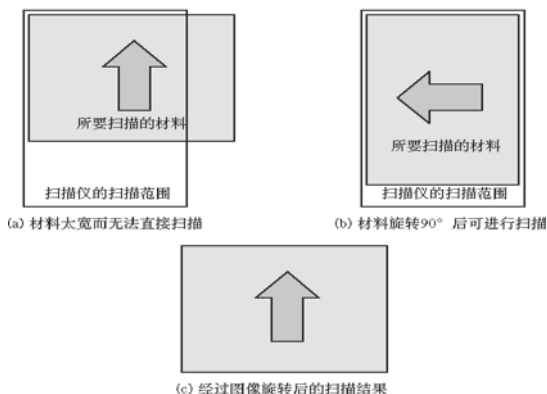


图 21 图像旋转功能的主要用途

3.2.3. 图像剪裁 [C]

选定【图像裁剪[C]】项，可以保留图像中所有选定的区域，去掉周围没有选定的区域。



图 22 (裁剪前)



图 23 (裁剪后)

为了方便起见，裁剪后的图像，仍然保留成矩形的。裁剪图像不是必须的，但利用它可以有效地节约图像的存储空间，提高处理速度。

3.2.4. 局部图像处理

局部图像处理包括【清除区域[I]】、【反转区域[R]】、【删除区域[D]】、【删除所有区域[A]】、【删除框线】、【删除所有框线】等六项命令。与整幅图像处理不同的是，局部图像处理的结果，只影响图像的当前区域，而不影响当前区域以外的图像。在计算机的屏幕上，当前区域有一个显著的标志，就是在其四个角上，都有一个小方块光标。整个区域显示为黄色。

当前区域有一个显著的标志，就是在其四个角上，都有一个小方块光标，且显示为黄色。

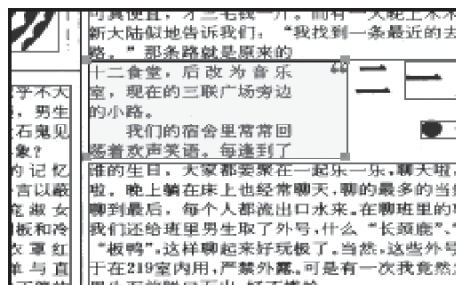


图 24

a) 清除区域 [L]

在【图像 [I]】菜单中，选定【清除区域 [L]】项或直接按空格键，当前区域的那部分图像就会被清除而变为空白。

在实际应用中，可能会遇到这样的情况：需要识别处理一幅图像的绝大部分，而只有图像中某一个部分是不需要的。这时将无用区域选作当前区域，再使用【清除区域 [L]】命令，就可以将无用区域清除掉。

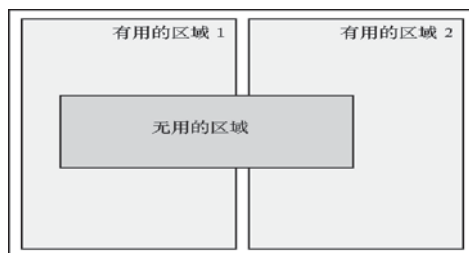


图 25 需要清除的区域示意图

b) 反转区域 [R]

对所选定的当前区域进行黑白反相。

c) 删除区域 [D]

在【图像 [I]】菜单中，选【删除区域 [D]】项或直接按键盘上的删除键 **Del**，可以删除确定当前区域的区域框，而区域中的内容保持不变。

d) 删除所有区域 [A]

删除所有区域的功能，与删除一个区域的功能非常类似，不同的是它一次就删除了划分区域的所有区域框，而不只是删除当前区域区域框。

e) 删除框线

在表格操作时，删除选定的一条表格线。

f) 删除所有框线

在表格操作时，删除所有表格线。

3.2.5. 恢复图像 [T]

用户如果决定放弃前面对图像所做的各种处理，可以选用此项命令，重新装入原图像。

3.2.6. 设置版面区域类型

TH-OCR 联想专业增强版系统的版面区域共有四种类型，即横排正文、竖排正文、表格和图形图像。在进行识别前必须正确定义图像中各区域的区域类型。

设置版面区域类型方法是：对于所划分的每一个当前区域，在【图像 [I]】菜单中选定相应的项目（【横排正文[H]】、【竖排正文[V]】、【表格[F]】、【图形图像[G]】）；或直接用鼠标按工具条中类型按键。

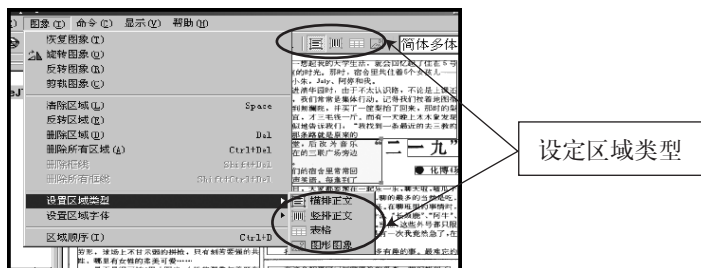


图 26

3.2.7. 区域顺序

执行这个命令时，则显示图像各个区域的序号。注意，当需要导出 TXT 文件格式时，须按照段落顺序手动调整区域顺序。

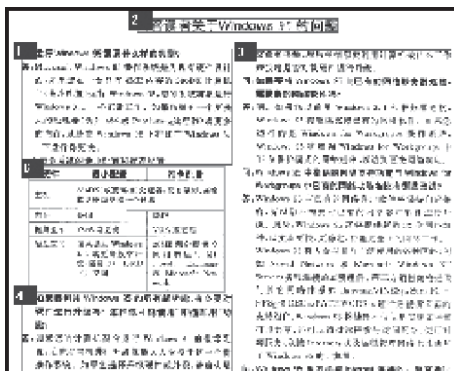


图 27

3.3 命令 [C]

【命令 [C]】的子菜单中共有 12 个选项。其中【设置 [S]】用于整个系统工作状况的调整，【倾斜校正 [W]】、【版面分析 [L]】、【框线检测 [F]】等类命令用于进一步完成识别前的准备，【全部识别 [A]】和【识别 [R]】命令为系统的核心命令，其它命令都服务于识别过程中及识别后的编辑及修改。【命令 [C]】的子菜单如图 28 所示。



图 28

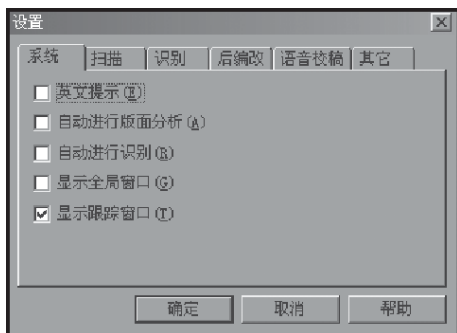


图 29

3.3.1. 设置 [S]

打开【命令[C]】菜单，选择【设置[S]】后，屏幕上显示【设置】对话框。如图所示，它包括了对系统、识别、扫描、后编改及其它等五设置的选择。

图 29 还表示了选择【系统】设置时的情况。【系统】设置包括【英文提示[E]】、【自动进行版面分析[A]】、【自动进行识别[R]】、【显示全局窗口[G]】、【显示跟踪窗口[T]】等项。注意：当选择【英文提示[E]】时，系统全部菜单均用英文表示；当选择【自动进行版面分析[A]】、【自动进行识别[R]】时，只要装入图像，系统就自动完成整个识别过程。建议对于复杂版面不选用此项。

【扫描】设置中包括【文件名】、【扫描图像格式】、【图像压缩格式】。图 30 用户可以根据实际工作情况自行选择。

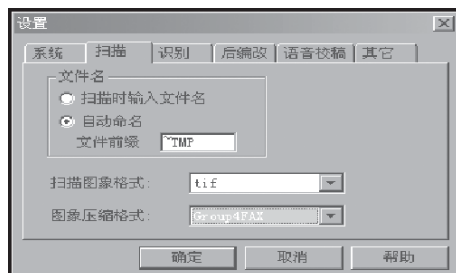


图 30

【识别】设置中包括【识别字体 [O]】、【输出全角字符 [F]】、【汉字后加空格 [V]】、【输出到剪贴版 [B]】、【输出到其它应用程序 [U]】。如图 31 用户可以根据自己的需要自行选择。

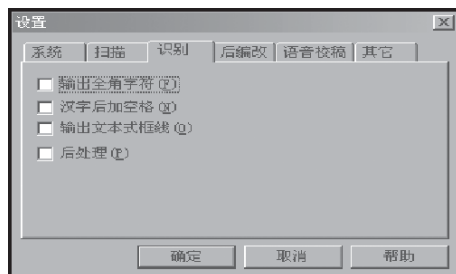


图 31

【后编改】设置中包括【浮动跟踪窗】与文本的位置关系、前景色、背景色及可疑字颜色，如图 32 所示，用户可根据自己的喜好设置。



图 32

【其它】设置中包括【表格[F]】中的【快速框线检测[Q]】与【输出表格线[L]】的选项，还包括【版面分析[A]】中的【杂志[M]】与【报纸[N]】选项，用户可以根据自己的需要自行选择。

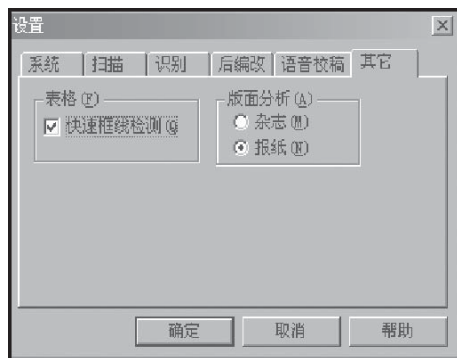


图 33

3.3.2. 倾斜校正

由于印刷或用户操作等多方面的原因,扫描得到的图像可能有一定的倾斜角度。尤其是小的倾斜角度,在扫描中难以避免。对于特别小的倾斜角度(约 $1^{\circ} \sim 2^{\circ}$,因具体情况的不同而异),TH-OCR 联想专业增强版系统可以自动适应,无需任何处理就能识别;而对于较大的倾斜角度,则需

要进行倾斜校正。一般当倾斜角度小于 $10^{\circ} \sim 15^{\circ}$ 时，可以先进行倾斜校正，然后再进行识别处理；如果倾斜角度超过 15° ，则由于倾斜校正会产生较大的失真和误差，从而影响识别结果，建议重新扫描图像。

系统提供自动及手动两种倾斜校正的方式。

a) 倾斜校正 [W] (自动倾斜校正)

执行【命令[C]】菜单中的【倾斜校正[W]】或用鼠标选工具条中的，系统会自动将倾斜的图像校正。

按压鼠标左键在倾斜的图像中拉出一块大小适当的区域，再选【命令[C]】菜单中【倾斜校正[W]】或用鼠标选工具条中的，系统会自动将倾斜的图像校正，如图 34 所示。

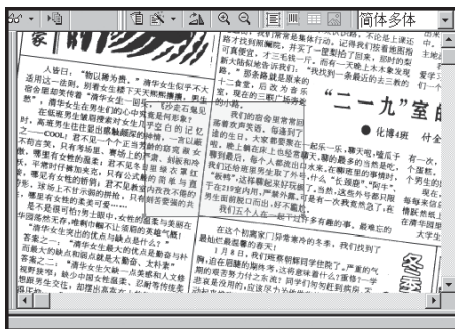


图 34

b) 手动倾斜校正 [D]

按住键盘上的 SHIFT 键，同时按鼠标右键并拖动使拉出的直线平行于倾斜的文本行，先松开右键后系统会自动将倾斜的图像校正，如图 35 所示。

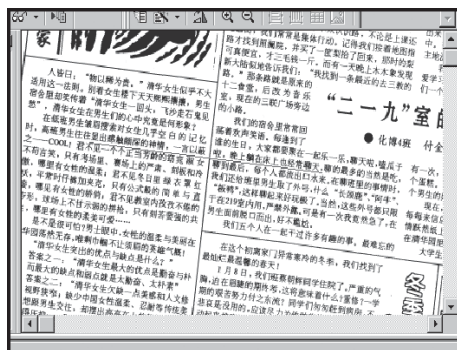


图 35

3.3.3. 版面分析

版面分析是将扫描得到的图像，划分出每一个区域块。对于各个不同的区域块，不仅给出其自身的属性（如横排的文字、竖排的文字、图形图像以及表格等等），而且标明不同区域块之间的顺序，以便系统进行识别处理。在 TH-OCR 联想专业增强版系统【命令[C]】菜单中，选【版面分析[L]】项（也可以直接按【F6】键），或用鼠标按工具条中快捷键，图像自动分为若干个属性不同的区域（图形图像、竖排文字、横排文字、表格等）。对于特别复杂的图像版面，自动分析后，须人工手动来辅助调整。

建立一个新的区域：在未被选定的区域，按住鼠标左键从鼠标所在位置（图中“+”符号处）拖画出一个方框将所需要的区域围住，如图 36 所示。

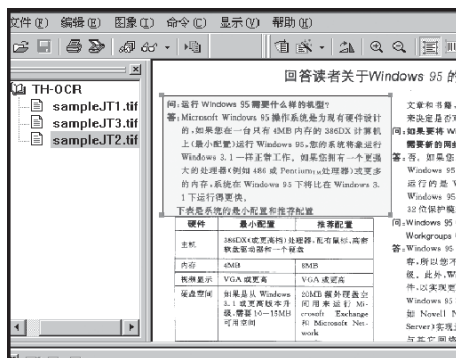


图 36

3.3.4. 框线检测 [F]、虚拟框线检测 [I]

TH-OCR 联想专业增强版系统无法识别未进行框线检测的表格图像, 也无法识别框线不全的表格图像。因此如果要识别的图像中含有表格, 必须先对它进行框线检测。

a) 框线检测 [F]

图像内如含表格, 应先将表格建为独立的区域, 然后确定其表格属性。这时系统即自动进行框线检测并将检出的框线 (包括内部的表格线) 用粉红色线表示。

b) 虚拟框线检测 [I]

当要识别的图像表格框线不全时, 必须进行虚拟框线检测。选择【命令 [C]】菜单的【虚拟框线检测 [I]】, 系统会将框线自动修补好。

注意: 虚拟框线检测必须在框线检测完成后才能进行。

3.3.5. 识别 [R]

图像文件根据各自情况经过倾斜校正、版面分析、框线检测及虚拟框线检测等处理后, 即可进入识别程序 (图像中只含有单栏的横排文字, 没有其它复杂的内容, 可以不进行版面分析, 就能够识别)。这是 TH-OCR 联想专业增强版系统的核心。为了保证正确的识别, 用户应按以下步骤进行:

a) 选择正确的识别字体

用鼠标打开工具条上的字体选择键 ，TH-OCR 联想专业增强版系统能识别的字体呈现在屏幕上，请用户根据被识别图像的具体情况选择字体：

简体多体（印刷体）——常见的宋体、仿宋体、黑体、楷体、圆体和魏碑等等。

繁体多体（印刷体）——常见的宋体、仿宋体、黑体、楷体、圆体和魏碑等等。

纯英文（印刷体）——常见的多体纯英文。

手写体——手写识别对文稿有一定的要求，须规范书写。

简体全字集——国标一二级全部字符集，魏碑、隶书等更多的字体。

繁体全字集——繁体一二级全部字符集。

b) 识别 [R]

在【命令 [C]】菜单中，选【识别 [R]】项（也可以直接按键盘上的【F2】键），或用鼠标按工具条的 ，系统即进入识别程序。识别过程中在识别输出窗口将出现一活动蓝条以报告识别完成的进度，如图37所示。

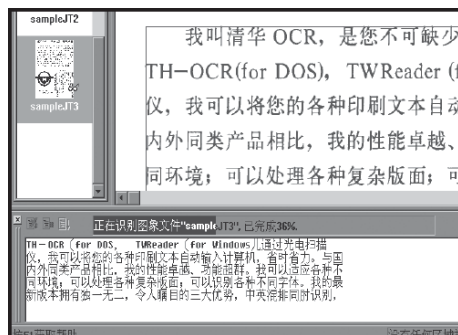


图 37

如果对已经识别处理过的图像再次选择识别命令，系统会显示出【覆盖】对话框，提示是否覆盖已有的识别结果，如图 38 所示，用户可以根据需要自行选择。

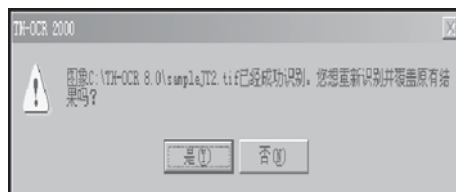


图 38

c) 识别所有页 [A]

当大量文稿需要录入时，可以把稿件逐页扫描，系统将在工程管理窗口内自动编入页码。在对每页文稿进行版面分析后，选【全部识别[A]】项（也可以直接按键盘上的【Ctrl +F2】），或用鼠标按工具条中的多页识别按钮。系统即进入批量识别程序。批量识别时的图像与识别时相似，此时右下角将报告已完成识别的图像的序号。

注意：使用批量识别功能时，必须保证每页图像都完成了倾斜校正、版面分析、框线检测等，同时每次页数不宜太多。

3.3.6 用户自学习 [U]

TH-OCR 联想专业增强版系统为用户提供了自学习功能。在识别中如果遇到生僻的汉字，甚至是某种特定的符号，都可以让计算机学习。这样以后再次遇到时，系统就能识别。

将光标移至需要学习的字符上，然后在【命令[C]】菜单中选【用户自学习[U]】，屏幕显示【用户自学习】对话框，如图 39。

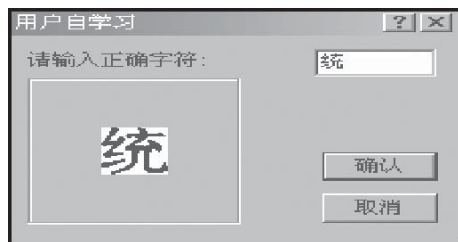


图 39

例如，上图对话框中左侧的“统”为当前汉字的图像，系统把它识别为“流”字，自学习时请在对话框右上角的编辑框中输入正确的汉字“统”，这样识别结果就更改过来了，如果需要还可以将输入的这个字送到用户字库中。

本系统可以学习 1024 个新的字符。但在使用该功能时应掌握以下原则：

- 1) 所要学习的字，其图像应是比较清楚的，否则会引起负效应，从而影响别的字的正确识别；
- 2) 学习的目的是为了应用，只有经常要用到的字符才适合于学习。

3.3.7. 修改用户库 [M]

当需要调整用户字库（比如想删去其中不再常用的字时），可以用【修改用户字库 [M]】命令。

在【命令 [C]】菜单中选【修改用户字库 [M]】项，屏幕显示【修改用户库】对话框，用鼠标选中所要删除的字（一个或多个均可，比如“统”字），经确认后再选【删除 [D]】或【全部删除 [A]】就可实现删除的目的，如图 40 所示。

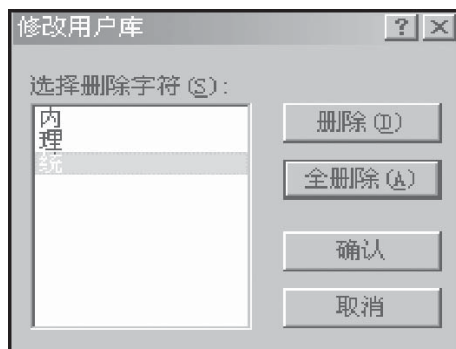


图 40

3.4 编辑 [E]

TH-OCR 联想专业增强版系统的编辑环境包括使用 Windows 标准的编辑操作和使用 TH-OCR 联想专业增强版系统特有的编辑功能两大部分。图 41 是【编辑 [E]】子菜单图，图 42 是在编辑环境下有关命令对应的工具条上的功能键图。

清除(C)	Ctrl+G
剪切(D)	Ctrl+X
复制(C)	Ctrl+C
粘贴(V)	Ctrl+V
清除(C)	F6
全选(A)	Ctrl+A
查找(F)	Ctrl+F
替换(R)	Ctrl+H
查找下一个(N)	F3
前一可换字符(B)	Ctrl+Up
后一可换字符(B)	Ctrl+Down
向前词汇(H)	Alt+F5
向后词汇(H)	Alt+F6
相似字(S)	Alt+F7
常用符号(O)	Alt+F8
行逆序(I)	

图 41

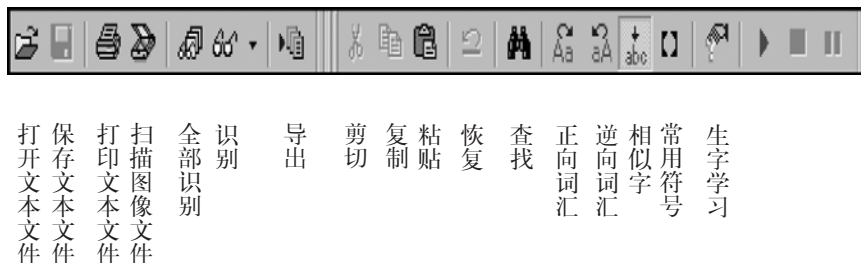


图 42 工具条上的编辑功能键

3.4.1. 在可疑字之间快速移动光标

TH-OCR 联想专业增强版系统不仅具有很高的识别正确率,而且对于没能正确识别的汉字,也可以非常方便地进行编辑修改。在实际的系统中,正常的文本是黑色的,可疑字是有特殊颜色的(颜色可以在【设置[S]】中选定),用户可以在这些有特殊颜色的可疑字之间快速移动光标。快速移动光标的方法是:同时按住【**Ctrl** + **↑**】键或【**Ctrl** + **↓**】键,或在【编辑[E]】菜单中选【前一可疑字符[V]】项或【后一可疑字符[X]】项。

3.4.2. Window 标准的编辑操作

Windows 标准的编辑操作,包括剪切、复制、粘贴和清除等四项操作。在 TH-OCR 联想专业增强版系统的【编辑[E]】菜单中,提供了这四项操作,主要面向文本进行处理。和所有具有编辑功能的 Windows 软件一样,TH-OCR 联想专业增强版系统的工具条中也设置了   , 分别实现剪切、复制和粘贴的操作。

3.4.3. TH-OCR 联想专业增强版系统特有的编辑功能

TH-OCR 联想专业增强版系统特有的编辑功能主要包括【前向词汇

[F]、**【逆向词汇 [B]**、**【相似字 [S]** 和 **【常用符号 [M]**，此外还包括虽不太常用但却很有用的 **【行逆序 [V]**。在 TH-OCR 联想专业增强版系统的编辑环境中，屏幕上既有识别结果文本，又有与之相对应的原始扫描图像，对应于识别结果文本中当前光标所在位置的汉字，图像中相应的字用一个蓝色方框包围。用户不必查阅原稿，就可进行全部的编辑校对和修改工作。

在 **【前向词汇 [F]**、**【逆向词汇 [B]**、**【相似字 [S]** 和 **【常用符号 [M]** 的编辑方式中，系统处理的都是当前光标所在位置的字符，以下各种描述也都对当前光标所在位置而言。

a) 前向词汇 [F]

前向词汇，就是根据光标所在位置的前一个汉字，以词汇联想的方式，提供出光标所在位置可能的汉字。采用前向词汇的编辑方法，是非常方便而有效的。比如文件材料中本来应是“我们”，而其中的“们”字由于种种原因，被计算机识别错了，在 **【前向词汇 [F]** 的方式下，系统会由于其前一个字是“我”而自动列出“校”、“爱”、“国”和“们”等一系列可以和“我”字组成词汇的汉字，只要简单地在其中选择“们”字即可完成编辑改错。

b) 逆向词汇 [B]

与前向词汇非常相似的另一种编辑修改方式，叫 **【逆向词汇 [B]**。逆向词汇，就是根据光标所在位置的后一个汉字，以词汇联想的方式，提供出光标所在位置可能的汉字。这也是非常方便而有效的。例如，文件材料中本来应是“清华”，而其中的“清”字，被计算机识别错了，在 **【逆向词汇 [B]** 的方式下，系统会由于其后一个字是“华”而自动列出“升”、“荣”、“清”和“年”等一系列可以和“华”字组成词汇的汉字，只要简单地在其中选择“清”字即可完成编辑改错。

c) 相似字 [S]

TH-OCR 联想专业增强版系统提供的第三种编辑修改方式，是 **【相似字 [S]**。汉字识别的过程，就是系统按照一定的算法和准则，确定出一系列与扫描得到的图像比较相似的汉字，其中最为相似的一个也就是识别

结果。而这里所说的相似字,则是指由系统列出所有这些可能的汉字,从而供用户选择正确的结果。在 TH-OCR 联想专业增强版系统的识别输入过程中,无论识别出错的汉字属于一个词汇,还是一个与前后汉字都无法组成词汇的单字,采用相似字的编辑方法,都是比较方便而有效的。因此,TH-OCR 联想专业增强版系统在进入编辑环境后,总是自动处于相似字状态,便于用户进行编辑修改。

【前向词汇[F]】、【逆向词汇[B]】和【相似字[S]】TH-OCR 联想专业版系统中的三种极为有效的编辑修改方式。但是,如果这三种方式中都没有所希望的汉字,则可以使用 Windows 环境下的任何一种汉字输入方法(如拼音、五笔等),进行修改。

d) 常用符号 [M]

TH-OCR 联想专业增强版系统还提供了【常用符号[M]】的编辑修改方式。常用符号,主要指那些键盘上不易输入而又常常用到的标点或其它符号。如果识别出错的是符号,就可方便地通过【常用符号[M]】的方式进行编辑修改。当切换到【常用符号[M]】状态后,屏幕左下角的编辑修改状态区域内,显示出“常用符号”四个字,并在其后面的编辑修改候选字区域内,列出常用的符号。屏幕上每次列出的常用符号都是十个,可以按键盘上的【Shift+>】键或【Shift+<】键进行滚动显示,也可以用鼠标按屏幕上符号前面的->或<-进行滚动显示。如果屏幕上显示出所需要的符号,按符号前面相应的数字键或直接用鼠标按所需的符号即可。

3.4.4. 行逆序 [V]

行逆序的功能是专门为港台的部分报纸和其它印刷品设计的。一般横排文字的排列是从左到右,而竖排文字则从右到左。然而在部分繁体材料中,有时会出现标题中的英文及数字从左到右排列,而横排汉字却从右到左的情况。当这样的材料被识别后,就必须调整识别结果中行内的文字顺序。这时使用行逆序就很方便了。行逆序后英文及数字顺序不变。

3.5 显示 [V]

图 43 为【显示 [V]】命令的子菜单内容。

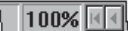


图 43

3.5.1. 工具条 [T]、状态行 [S]

工具条和状态行分别位于屏幕的上边和下边。当这两项被选中时,【显示 [V]】菜单中对应项前有“✓”; 用户如果想取消【工具条 [T] 或【状态行 [S]】,可在【显示 [V]】的子菜单中选中对应项,使所选项前面的“✓”消除。

3.5.2. 局部图像的显示比例

局部图像的显示比例是可以改变的,其操作十分简单方便,而且局部图像的当前显示比例,在状态行中的图像显示比例区域  中可以看到。

a) 放大 [I]: 在【显示 [V]】的子菜单中选择【放大 [I]】项,或直接按加号键 , 也可以用鼠标按工具条中的放大显示键 , 局部图像

逐渐放大，显示比例依次为 10 % → 25 % → 50 % → 75 % → 100 % → 200 %；

b) 缩小 [O]: 在【显示 [V]】菜单中选择【缩小 [O]】项，或直接按减号键 [$\square$], 也可以用鼠标按工具条中的缩小显示键 ，局部图像逐渐缩小，显示比例依次为 200 % → 100 % → 75 % → 50 % → 25 % → 10 %。

c) 选择比例 [Z]: 当选择了这一项时，屏幕上显示【缩放比例】对话框。

如图 44 所示，对话框中列出了所有的显示比例及【窗口大小 [W]】等七个选项。

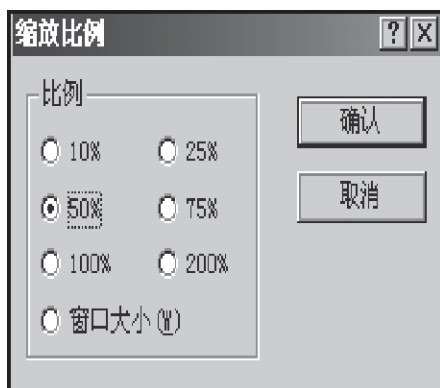


图 44

当选择了【窗口大小 [W]】选项时，全局图像窗口和局部图像窗口中显示的都是整个图像。此时工具条中的放大、缩小都变为无效状 ；同时，状态行中的图像显示比例区域显示出自动二字 。如果在选了【窗口大小 [W]】项后，又希望重新改变局部图像的显示比例，则必须通过【显示 [V]】菜单中的【选择比例 [Z]】项，重新进入【缩放比例】对话框进行改变。

3.6 帮助 [H]

无论是在图像环境中，还是在编辑环境中，用鼠标按工具条中的【关于】按键，都会显示出一个“关于TH-OCR联想专业增强版For Windows”对话框。而有关【帮助[H]】菜单的操作，均是Windows本身所有的，用户可参考有关资料。

3.7 退出系统

在TH-OCR 联想专业增强版系统的图像环境下，从【文件[F]】菜单中选【退出[X]】项或直接按键盘上的【+】键，也可以从系统菜单『』中选择【关闭[C]】项，均能关闭所有的文件退出TH-OCR 联想专业增强版系统。



产品名称：TH-OCR联想专业版

公司名称：联想（北京）有限公司

联系地址：北京市海淀区上地信息产业基地创业路6号

邮政编码：100085

网址：www.legend.com

版本号：2.0版

本说明书版权所有（C）（2002）联想（北京）有限公司

LJ7190001