紫光文字识别软件

一、软件说明

紫光以清华大学电子工程系为技术依托，隆重推出“基于识别的原文重现”自动电子出版物制作系统（TH-OCR 紫光专业版）。该系统通过了国家教育部组织的专家鉴定会的鉴定。该出版系统是国内外首次推出的能同时识别超大字符集（13051 字）和超多种汉字字体（近百种）、并且将电子文档的错误率降低到万分之一以下的、能将复杂报纸杂志文档经版面分析、识别、理解，最后自动精确重构为原式原样的标准格式电子文档的电子出版系统，为我国信息资源建设提供了一个快捷、高效的系统解决方案，是一个具有划时代意义的创举。TH-OCR 紫光专业版是一套理想的中英文印刷体自动识别系统，可广泛应用于办公自动化的资料录人、文献建档、资料处理、信息管理、智能翻译等领域。

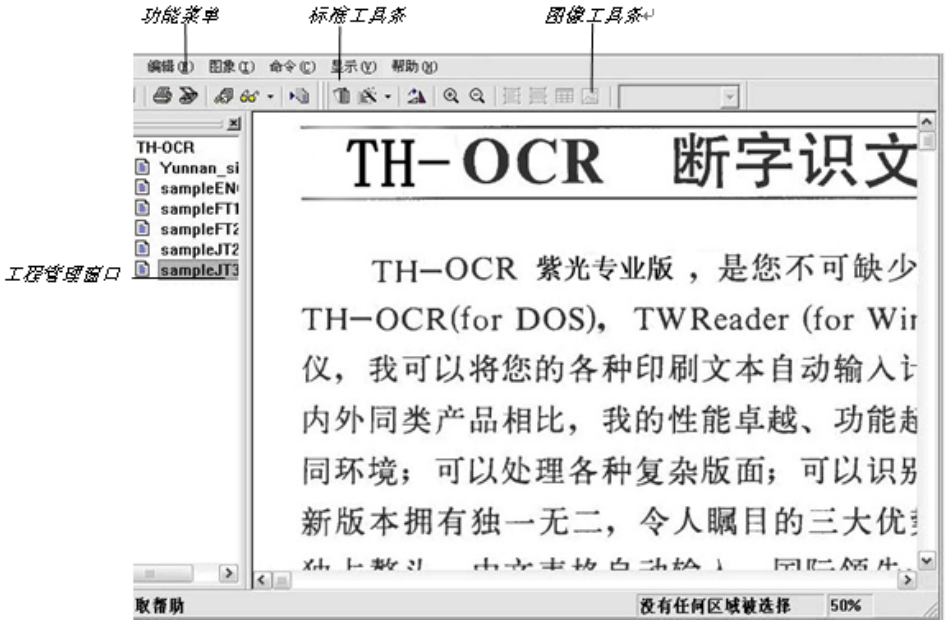
二、软件操作

注意：随紫光扫描仪一起赠送的 TH-OCR 紫光专业版，需要与紫光扫描仪配套使用，请在正确安装紫光扫描仪之后使用。

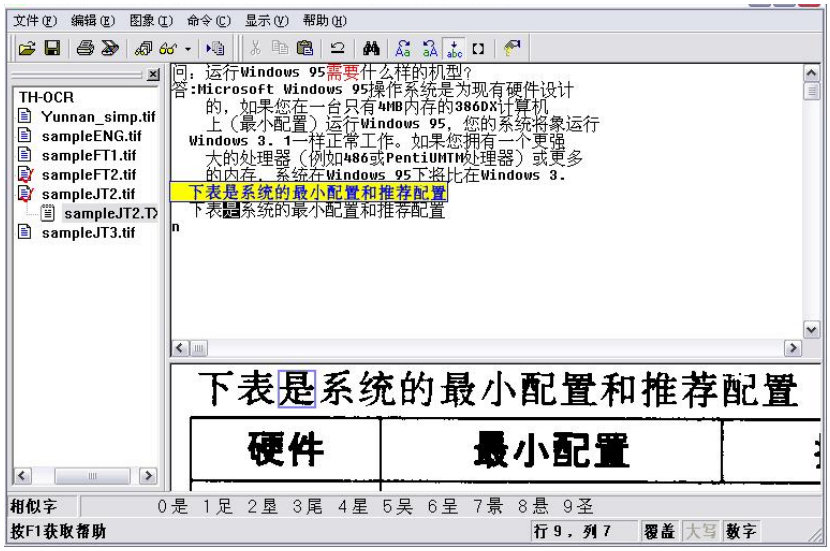
1、软件界面及功能说明

在不同的操作状态，分别有图像版面分析和编辑修改两种界面。

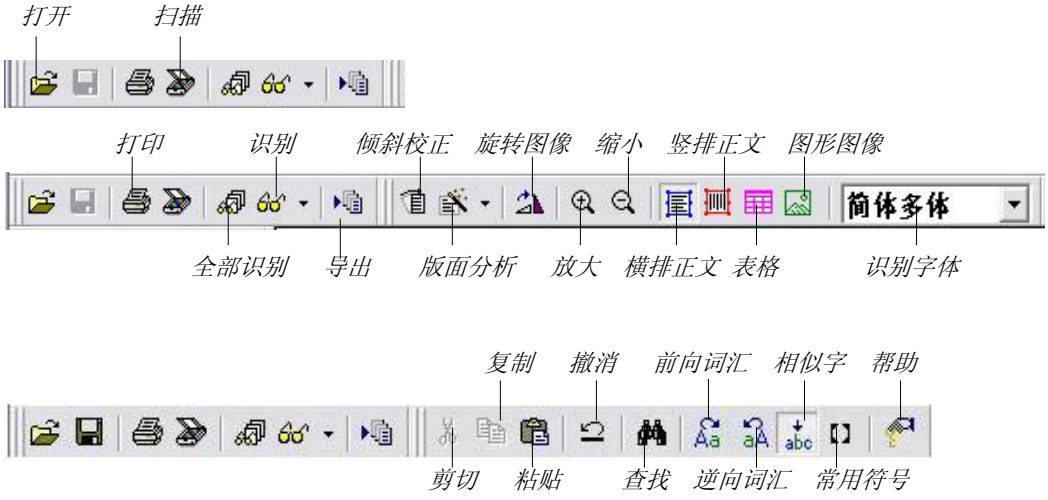
软件界面一：图像版面分析界面


功能菜单
标准工具条
图像工具条
工程管理窗口
编辑(E) 图象(I) 命令(C) 显示(V) 帮助(H)
TH-OCR
Yunnan_si
sampleEN
sampleFT1
sampleFT2
sampleJT2
sampleJT3
TH-OCR 断字识文
TH-OCR 紫光专业版，是您不可缺少
TH-OCR(for DOS)，TWReader (for Wir
仪，我可以将您的各种印刷文本自动输入计
内外同类产品相比，我的性能卓越、功能超
同环境；可以处理各种复杂版面；可以识别
新版本拥有独一无二，令人瞩目的三大优
取帮助
没有任何区域被选择
50%

软件界面二：编辑修改界面后改编工具条

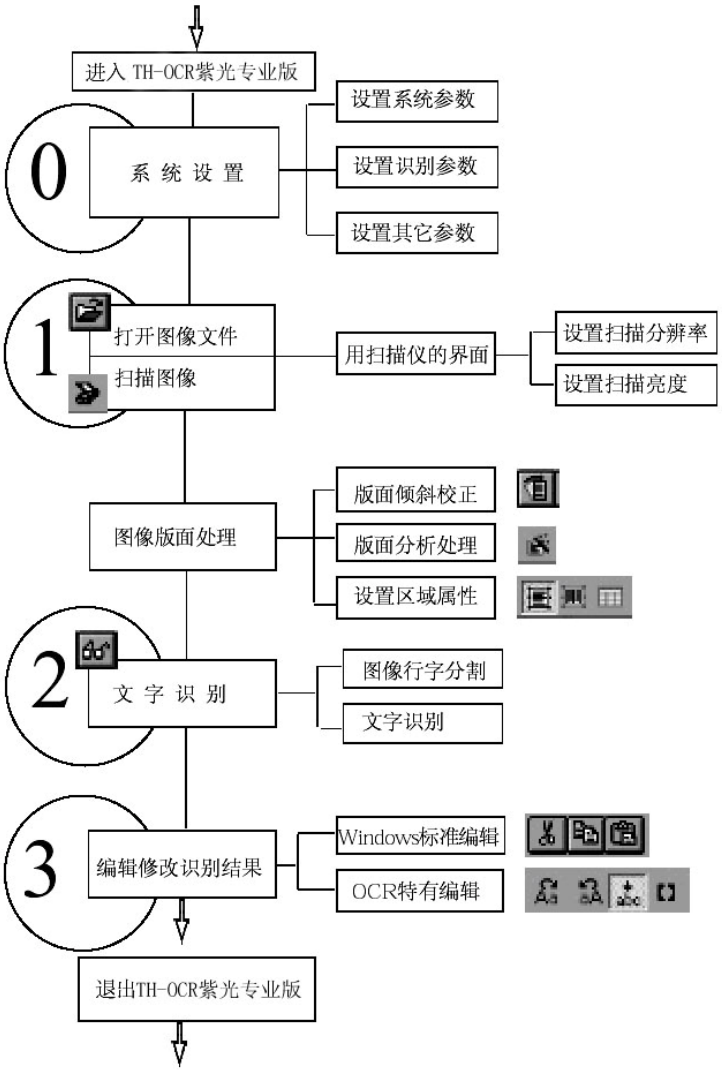


工具条中各快捷作按钮的功能说明如下：



2、操作流程

TH-OCR 紫光专业版的操作流程分为设置、获取图像、版面分析、文字识别、编辑修改等五步，如图所示：



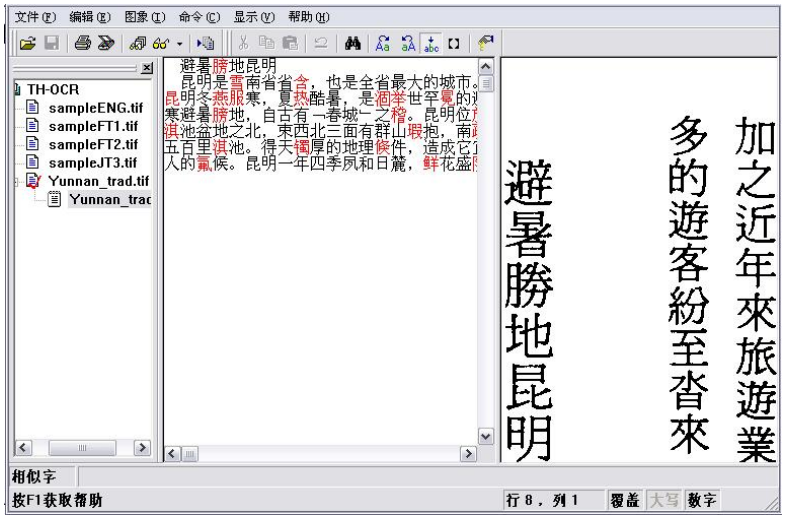
设置

使用系统前应根据应用环境及需求设定系统参数。从“命令”菜单选择“设置”命令，在打开的“设置”对话框中可对系统、扫描、识别、后编改及其它（表格、版面分析等）参数进行设定。这些参数一般按默认设置即可。

获取所要识别的图像文件

获取所要识别的图像文件有两种方式：通过扫描仪扫描新的图像或打开已有图像文件。

如果要扫描新的图像，则应从“文件”菜单选择“扫描”命令或单击工具条上的扫描图标，对所要识别的稿件进行扫描。扫描完成后退出扫描界面。识别后，文件将不直接出现在识别界面中，双击被识别图像文件左边的“+”号，出现识别后的文件名，双击该文件名，即可打开该文件进行编辑修改，当文件为横排时，右侧上部为识别后文本窗口，右侧下部为局部图像窗口；当文件为竖排时，识别后文本窗口和识别前局部图像窗口左右排列，如下图：



对于磁盘上原有保存好的图像文件，可以直接从“文件”菜单中选择“打开”命令或单击工具条上的打开图标，在“打开”对话框中指定路径、文件类型、文件名，单击“打开”按钮，即可将选定的图像文件显示在工程管理窗口。同 Windows 的操作一样，如要一次打开多个文件，请使用“Ctrl”或“Shift”按键选择相应文件打开即可。图像文件打开后，显示在图像窗口中，与扫描得到的图像相似。打开多个文件进行识别时，所识别得到的文本作为工程将按选中文件列表中的文件顺序排列在工程管理窗口。

提示：扫描仪的操作请参照随同扫描仪附送的用户手册的指导。

注意：

1. TH-OCR 紫光专业版 可以识别彩色、灰度和黑白二值的图像，扫描时的扫描模式可任意设置。
2. 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 300dpi，扫描亮度为自动或默认值；

3. 对已有的图像文件，要注意其图像存储格式是否符合 TH-OCR 紫光专业版 系统的要求(非压缩 TIFF 格式、PackBit 或 G4 压缩的 TIFF 格式、BMP 格式或 PCX 格式）。

进行图像版面处理

对扫描所得图像文件根据需要进行处理（旋转、反转、剪裁、倾斜校正等）和版面分析等，并选择需识别的内码、字体，为识别做好准备。

注意：

对于比较简单的结构，可以使用自动版面分析，如果版面较复杂，请手工进行版面分析，只需简单地用鼠标框选各识别区域，并根据原稿的版式选择要进行文字识别的区域的属性，如：横排正文、竖排正文、表格、图形等。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图形。

从“命令”菜单选择“识别”命令或单击工具条的识别按钮，完成版面的识别。完成后，双击被识别图像文件之后，再双击识别后的文件名，进人编辑修改状态。

在编辑修改状态，对于正常识别的文字用黑色显示，对于可疑字用系统设置中指定的颜色显示（默认为红色），便于提示修改。

系统提供 Windows 标准的编辑操作和 TH-OCR 紫光专业版 系统特有的编辑功能。Windows 标准的编辑操作包括剪切、复制、粘贴和清除等，TH-OCR 紫光专业版系统特有的编辑功能主要包括前向词汇、逆向词汇、相似字、常用符号、和行逆序。

通过双击项目管理窗口的原图形文件和被识别后文件，可以在图像状态和编辑状态间切换。

识别结果输出

识别结果经修改编辑后，可根据需要输出。

● 存盘：从“文件”菜单选择“另存为”命令，在“另存为”对话框中指定文件名后将文本以其它文件名保存。

● 导出：从“文件”菜单选择“导出”命令，在“导出为”对话框中指定文件名后可将识别后的稿件以包含版面格式的富文本格式（RTF）文件、包含

版面格式的页面格式（html 格式，可用 IE 5 等应用程序打开）以及纯文本的保存。

● 打印：建议通过Word 或IE 等编辑软件进行编辑后打印或按  ，进行打印。

退出

在识别过程中系统会生成跟踪文件。为节省硬盘空间，退出系统时，系统会将工作目录中“*.chr、*.sim、*.trc、*.rgn”等跟踪文件自动删除。

3、表格的识别与导出

对表格图像的识别参照如下步骤：

版面分析

1. 将表头或独立于表格的文本部分单独框出，并定义为“正文”属性，框线为兰色。
2. 将完整的表格图像框出，定义成“表格”属性，框线为粉红色。
3. 依原稿类型定义文字属性为简体多体、繁体多体或其它。

识别

对所选区域进行识别，识别完成后请进人编辑界面，可进行文字的编辑修改。

导出

由于包含表格，因此识别结果一定要导出为 RTF 或 HTML 格式，否则表格线是断开的。

提示：横排正文的框线为蓝色；竖排正文的框线为红色；表格的框线为粉色；图形的框线为绿色，图形框线内的内容是不被识别的。

4、倾斜校正

扫描时，原稿一定要摆放端正，若稍有倾斜可使用倾斜校正功能自动校正，若倾斜角度较大时，则需进行手动的倾斜校正。手动倾斜校正的方法是：按住 Shift 键的同时按鼠标的右键在图像中拉一条平行于倾斜文字行的直线，然后放开鼠标的右键，则校正完成。但如果倾斜角度太大（超过 15° ），则会

由于倾斜校正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。

5、调整文本顺序

在版面分析中设定多个文本识别区域时，识别结果将按区域的编号顺序排列。如需要调整识别区域的顺序，请在当前的区域内部，按鼠标右键，选择“区域顺序”，在每个选择区域的左上角显示该区域的序号，双击该序号即可调节，调节到你需要的顺序后，请在区域外的任何一处点击一下，则调节顺序被确认。

注意：

调节任何一个区域的序号后，其它相关区域的顺序也会相应改变。

6、导出单页或多页文本文件

对扫描或打开的图像文件识别后的文本文件，如选择“文件”菜单中的“导出”命令或直接调用工具条上的“ ”图标，会出现一个导出设置对话框，在该对话框中有“导出当前页”和“所有页导出为一个文件”两个选项，如选则“导出当前页”项，则只导出当前打开页的文件，如选择“所有页导出为一个文件”则将所有打开页的文件全部导出后合并为指定的一个结果文件，对于导出的文件，可以指定文件名、文件格式及文件的存储位置。

7、分辨率设定与字号大小对照表

扫描图像时分辨率的设定请根据文稿上文字的大小，参照下表的推荐值：

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	

8、常见问题

扫描时提示“装入 TWAIN.DLL 错误”。

请正确安装紫光扫描仪的驱动程序，连接好扫描仪，并将紫光扫描仪打开。

识别完成后屏幕为空白，只有光标闪动。

如果原稿中有图形，OCR 会认为此文件不符合要求而不作识别。此时应先进行版面分析，将所要识别的文字区域按顺序框出识别区域后再进行识别。

识别出的文字出现乱码。

1. 是否文字的方向不对，请正确调整文字方向。
2. 是否定义的文字属性（简体多体、繁体多体、纯英文、手写体等）与原稿不符，请设定相应的文字属性。
3. 是否原稿中的文字旁有辅助线，字体为斜体或艺术字等，此类原稿不能被正确识别。
4. 扫描时设置的分辨率是否不合适，请在扫描时参照分辨率设定与字号大小对照表中的推荐值选择适合的分辨率。
5. 扫描文稿时设定了镜像处理功能，扫描结果图像与原稿左右相反。
6. 原稿不清晰（如传真件、油印试卷、报纸等），若是报纸，可以适当地调节图像的对比度或亮度以得到较好的扫描效果，提高识别率。文章始部分识别率较高，但后面部分识别率低。
7. 原稿在扫描时摆放不正，若倾斜角度不大可进行倾斜校正，否则需重新扫描。

表格识别时，只识别出表头而无表格。

没有单独定义出表格属性。请按表格的识别与导出部分的说明进行版面分析。

识别繁体字得到简体字而非繁体字。

请从 Windows 操作系统的“开始”菜单指向“程序”中的“TH-OCR 紫光专业版”程序组，选中“选择系统内码”选项，在“Select System Inner－code”对话框中选定“GBK －code（All China）”。

如何使文件导出到 Microsoft Word 中。

导出时，选择“保存类型”为 rtf 即*.rtf，即可在 word 中打开。

如何使文件导出到 Microsoft EXCEL 中。

导出时，选择“保存类型”为 html 即*.htm，即可在 EXCEL 中打开。

如何使识别后的文件成为主页。

导出时，选择“保存类型”为 html 即*.htm 即可，同时保持原稿原版面。

9、取得帮助

从“帮助”菜单选择“目录和索引”命令，显示帮助主题，可以选择帮助主题或索引帮助内容；选择“快速指南”命令，显示 OCR 识别的基本操作流程。

注意： TH-OCR 紫光专业版主要用于识别简体中文，我们建议您在进行英语或其它欧洲语言的识别时，选用 ABBYY FineReader 文字识别软件。若您已由随机附赠的安装光盘安装此软件，您可单击 Windows 系统“开始”>“程序”来打开相应的识别软件。