

丹青

中英日文文件辨识系统

4.5 版



S/N:107450-02-01-W-Sc-070502-01

版权

版权所有：力新国际科技股份有限公司

初版日期：2002 年 8 月

本书版权为力新国际科技股份有限公司所有，未取得书面授权，不得将本书内容以任何形式复制、翻印，或以电子文件方式保存、运输。

程序光盘中的范例图片，仅供个人展示、制作及简报使用，不得用于商业拷贝、销售或流传等其他用途。

力新国际科技股份有限公司保留随时更新本书内容的权利。

商标

IBM PC 是 International Business Machines Corporation 的注册商标；MS Windows 98/Me/2000/XP 是 Microsoft Corporation 的注册商标；Pentium、MMX 是 Intel Corporation 的注册商标。

本书提及的产品名称皆为其所属公司的注册或未注册商标。

www.newsoftinc.com

www.newsoft.com.tw

www.newsoft.co.jp

de.newsoft.eu.com

newsoft.net.cn

目录

第1章

介绍	1
丹青的功能与特色	1
系统需求	4
硬件需求	4
软件需求	4
系统安装	4

第2章

基本概念	6
操作流程图	6
认识丹青界面	8
原稿图片模式:	8
全页图文模式:	11
文稿编辑模式:	14
改变屏幕配置	16
缩放图片显示比例	16
页面移动及信息查询	16
使用在线帮助	17
设定系统默认值	17

第3章

输入图片	20
扫描与打开	20
图片处理	21
旋转图片	21
清除杂点及补漏白	22
切除	23
反白功能	23

第4章

辨识文件	24
设定辨识字集	24
设定辨识区块	25
设定版面格式	25
栏位设定	25
排列设定	26
表格设定	26
内容设定	26
设定版面模板	26
版面分析	27
自动分析版面	27
手动设定版面	28
改变辨识顺序	28
保存版面	28
辨识文件	28
选择校对词库	30
设定校对词库	30
使用词库校对	31
自动辨识文件	31
自动辨识	31
设定自动模板	32

第5章

文稿校对	34
放弃辨识	34
校对文稿	34
再辨识	36
擦除杂点再辨识	36
修补图片再辨识	37
文字切割再辨识	37
文字合并再辨识	37
文字行切割再辨识	38
文字行合并再辨识	38

区块再辨识.....	38
区块结合再辨识	39
区块分开再辨识	39
学习新字	39
删除学习字	40
学习字设定	40

第 6 章

输出文件	42
保存辨识前的图片	42
保存辨识后的图文和表格	43
保存常用的版面格式	44
打印	45
发送	45

第 7 章

图文辨识范例	47
辨识含有图形及文字的文件	47

第 8 章

英文辨识范例	51
辨识英文文件	51

第 9 章

表格辨识范例	55
辨识一般表格图片	55
辨识暗线表格图片	59

第 10 章

自动辨识范例	64
自动辨识文件	64

附录 A	
用语说明	68
附录 B	
菜单	70
附录 C	
工具栏及其下拉式菜单.....	74
附录 D	
编辑工具箱图标	76
附录 E	
扫描的建议.....	79
如何改善辨识品质	79
图例一：扫描解析度建议	80
图例二：标准	81
图例三：太浓	81
图例四：太淡	82

第 1 章

介绍

丹青中英日文文件辨识系统能让您轻松且快速地将图形文件转换成可编辑的文本文件。其所能辨识的内容包括繁体中文、简体中文、日文、英文、阿拉伯数字及含表格的文件。辨识后的文本文件所占的内存空间远较未辨识前的图形文件小。在经过校对后即可保存成 TXT、RTF、DOC、XLS、SLK、CSV 等文件，并且可以在一般的文字处理软件中打开和编辑。您还可以将文件存成 HTML 格式，通过网络浏览器直接打开。

丹青的功能与特色

➤ 自动辨识

只要按一个按钮，便可自动分析、辨识、校对图片，并可转换成可编辑的文本文件。

➤ 高辨识速率

在一般的个人电脑上(Pentium III 667)，丹青系统每秒钟所能辨识的中文字高达 150 个。

➤ 可辨识中文繁体、简体、日文及纯英文文件

丹青采用多辨识引擎结构，可依据您的需要，辨识中文繁体、简体、日文及纯英文文件。

➤ 可处理多页文稿

可一次处理多达 50 页的文稿辨识。

➤ 可辨识黑白、彩色的文件

无论是黑白或彩色的文件，皆能获得极佳的辨识效果。

➤ 多字体辨识及重现

能够辨识多种印刷字体，如宋体、黑体、仿宋体、楷书、圆体、隶书等，并能在辨识后还原成原稿的字体。

➤ 原文重现编辑环境

辨识结果依照原文件的版面格式呈现，方便您校对编辑并节省重新排版的时间。

➤ 自动图文分离及文件再编辑

自动分离图片上的图形和文字，并可将辨识后的结果保存成 TXT、DOC、RTF 等文本文件，在一般文字处理器如 Word 中再编辑。

➤ 表格辨识功能

丹青能够辨识各种表格图片，并可将结果保存成 XLS、SLK、CSV 等文件在 Excel 中做进一步的处理，或将文件保存成 RTF 格式，在诸如 Word 之类的文字排版软件内重现原图片上的表格原貌。

➤ 可保存多种文件格式

辨识后的结果可保存成不同的文件格式，如 TXT、RTF、DOC、XLS、SLK、CSV、HTML 等，方便您做不同的应用与处理。

➤ 直接发送

可将辨识结果直接发送至您所指定的应用软件中再处理，例如可设定辨识结果自动保存成 HTML 文件并直接发送至网络浏览器中。

➤ 自动侦测图片倾斜角度

自动侦测图片倾斜角度，并提供旋转图片的功能。

➤ 自动校对

利用内建的常用词库自动校对辨识出的文字，并标示出辨识时所碰到的疑问字，节省您校对的时间。

➤ 候选字功能

提供字形上相似，或语意上前后相连的候选字，可供您轻松地更正辨识错误的文字。

➤ 学习新字功能

提供学习新字的功能，可将较易辨识错误的字元输入到学习资料库中，于下一次辨识时使用，以提高辨识的正确率。

➤ 资料交换

可经由中文 Windows 环境的剪贴板输入图片或是输出图文文件，与其它 Windows 应用软件交换资料。

➤ 中文横竖排及单多栏辨识

自动分析横排、竖排、横竖排并存以及单栏或多栏的图片。若输入的图片版面过于复杂，您也可以自己设定文字图片的格式，以利于系统做正确的分割及辨识。

➤ 存取版面

提供版面存取功能，方便您将常用的版面格式保存成丹青的版面文件 (*.tpl)，并可套用在新输入的文件图片上，以节省版面分析的时间。

➤ 文稿修改

提供图文对照(也就是原文字图片与辨识出的文字相互对照)的文稿校对功能，让您能够利用键盘、候选字或个人词库等更正辨识错误的文字。

➤ 再辨识功能

提供合/分字、合/分行、合/分区块、变更区块属性等再辨识功能，让您

能够修正错误的辨识结果，以便利校对文稿。

系统需求

硬件需求

- Pentium II (含) 以上的 IBM 电脑或其他相容电脑
- 64MB 以上内存(建议 128MB 以上)
- 屏幕解析度 800×600 Hi-Color
- 磁盘空间 300MB 以上
- 光驱
- 支援 TWAIN 界面的扫描仪（请参考扫描仪的使用手册）

软件需求

- Microsoft Windows 98/Me/2000/XP

注意: Windows 98/Me 不支持“多语种用户界面设置”功能。

系统安装

1. 将丹青的安装光盘放入光驱中。
2. 依照屏幕上的指示完成安装程序。

注意: 在 Windows 2000/XP 环境下安装丹青系统前，请先将工作环境恢复至计算机默认的语言界面，以避免在安装过程中出错。

安装完成后，点选“开始”菜单中的“程序”，选择“丹青中英日文文件辨识系统>丹青中英日文文件辨识系统”，开始运行本软件。您也可以双击桌面上的丹青图标，直接进入丹青系统。

注意: 在安装完成后, “丹青中英日文文件辨识系统” 的选项组中会包含以下选项:

“多语种用户界面设置”: 您可使用此功能切换丹青系统至下列的设置: 繁体中文、简体中文、日文、自动检测。请选择[开始]菜单中的[程序], 选择[丹青中英日文文件识别系统>多语种用户界面设置], 设置语言后, 打开丹青程序, 程序将会切换至您所选择的语言使用界面。(此功能须计算机工作环境可支持多语言)。

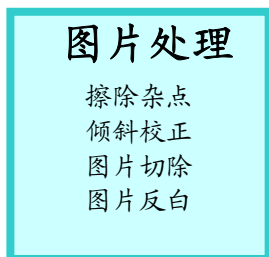
“卸载丹青中英日文文件辨识系统”: 您可以执行这个程序, 完整地卸载丹青系统。

第2章

基本概念

本章为您描述使用丹青中英日文文件辨识系统时，必备的基本概念及操作技巧，如：了解丹青系统操作流程及操作界面、改变屏幕配置、缩放图片显示比例、查看文件信息、在不同页面间移动、设定系统默认值，以及使用在线帮助等等。

操作流程图



辨识文件

自动辨识
图文辨识
表格辨识
英文辨识
简繁体字辨识



校对文稿

字词校对
再辨识
学习新字



输出文件

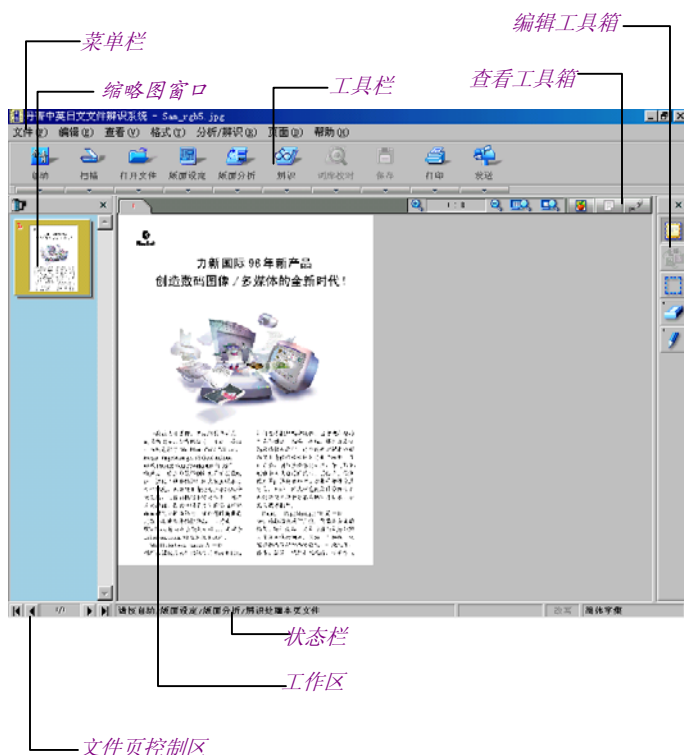
保存图片
保存文本文件
打印
发送

认识丹青界面

丹青系统包含三种操作界面模式：

原稿图片模式：

在原稿图片模式中，您可以完成辨识前的所有准备工作，如通过扫描仪或磁盘驱动器输入图片，使用编辑工具修饰图片，以及最重要的辨识项目设定（如设定辨识语言、文字排列的方式和是否含有表格等）。



原稿图片模式

菜单栏

菜单栏分类列出所有可供您使用的命令，如打开、保存、编辑或辨识文件等。

文件(F) 编辑(E) 查看(V) 格式(T) 分析/辨识(R) 页面(Q) 帮助(H)

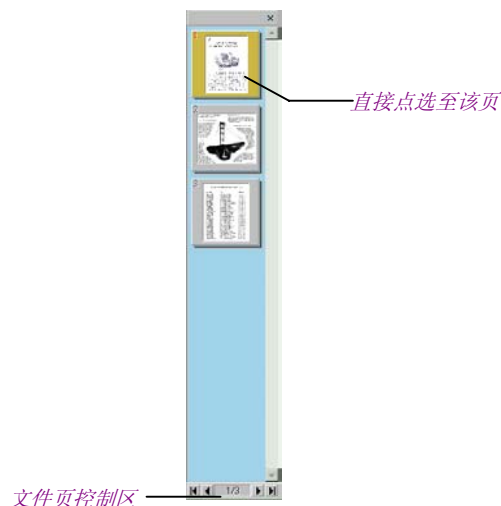
若您要打开一个菜单，您只要在菜单上按一下鼠标左键或是按一下该菜单上提示的英文字并同时按住“Alt”键即可打开。

同样地，如果您要执行某个命令，您只要将光标移到该命令上然后按一下鼠标左键或是按一下该命令提示的英文字即可。

若您不想执行任何命令而要关闭已打开的菜单，您只要在该菜单外任意处按一下鼠标左键或按“Esc”键即可。

缩略图窗口和文件页控制区

缩略图窗口将已打开的文件以缩略图方式显示，可供您直接选取。您也可以利用下方的文件页控制区的按键，移动至您想要的页面。



工具栏

工具栏上的图标可以让您快速且轻松地执行各种菜单命令；每个工具栏下方都有一个下拉式的命令菜单，可提供您更多的选择。



按此处调出下拉式菜单

工作区

您可在工作区中执行图片处理的工作，例如转正倾斜的文件图片，清除杂点，切除不需辨识的部分，使得辨识结果更令人满意。

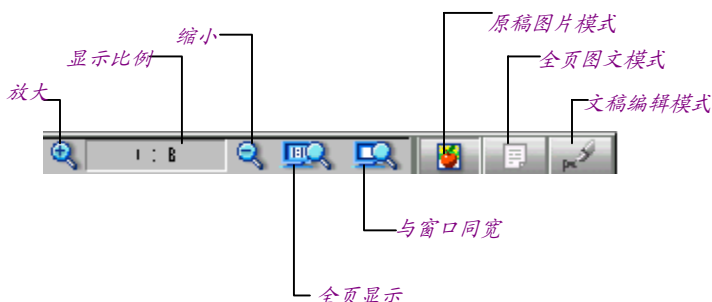
状态栏

状态栏显示目前光标所在位置的 X 座标和 Y 座标、光标所在对象的相关信息、文字输入模式，以及目前所选择的辨识字集，可点击状态栏中的右方选择要使用的识别词汇表。



查看工具箱

查看工具箱可让您缩小或放大图片的显示比例，并可选择原稿图片模式、全页图文模式及文稿编辑模式。



编辑工具箱

您可以利用提供的编辑工具，处理辨识过程中各个阶段的文件，例如编辑扫描进来的文件图片稿、更改文件的区块设定、校正辨识后的文稿等等。编辑工具箱也会随着辨识阶段的不同提供您不同的编辑工具。以下为版面分析前后原稿图片窗口所分别显示的编辑工具箱。



版面分析前

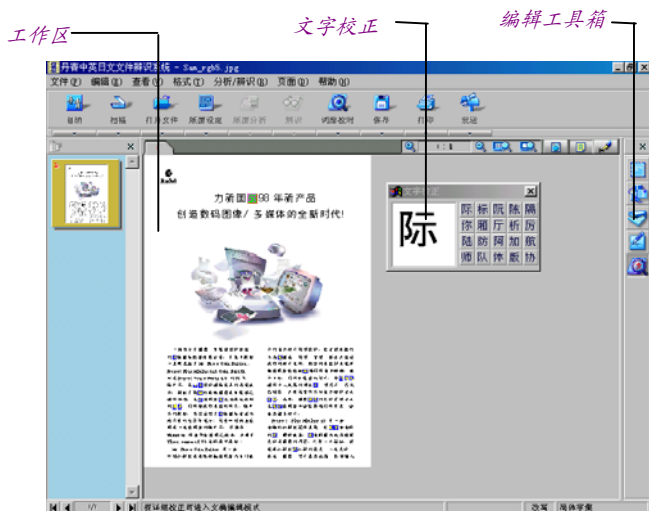


版面分析后

原稿图片窗口在版面分析前后所分别显示的编辑工具箱



全页图文模式：

“全页图文模式”可供您查看辨识后的文件全貌，并调整与版面相关的设定，如合并/分割区块、更改区块属性、调整区块的辨识顺序等；您也可以在此模式中直接校正辨识错的字。




全页图文模式

工作区


工作区内显示辨识后的文件全貌。若您按下文字校对工具，工具区内会出现一些蓝底黄字的字样，其代表的是系统在辨识后所显示的疑问字；若您按下变更辨识区块顺序工具，各区块会以红线框出，并标示出区块的辨识顺序编号。

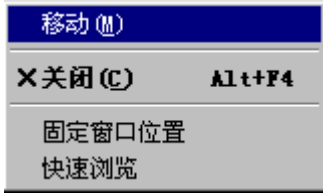
文字校正窗口

当您发现系统辨识错误的字时，可选择“文字校对工具”，并在该错误字上按一下，工具区内将会出现如下的窗口：



文字校正窗口

您可以按一下，在菜单中设定文字校正窗口属性：



文字校正窗口的菜单

- 移动** 选择以键盘的上下左右键移动文字校正窗口的位置。
- 关闭** 关闭文字校正窗口。
- 固定窗口位置** 在您用鼠标拖动文字校正窗口至合适的位置后，可选择此项以固定窗口位置；若您没有选择此项，窗口位置则

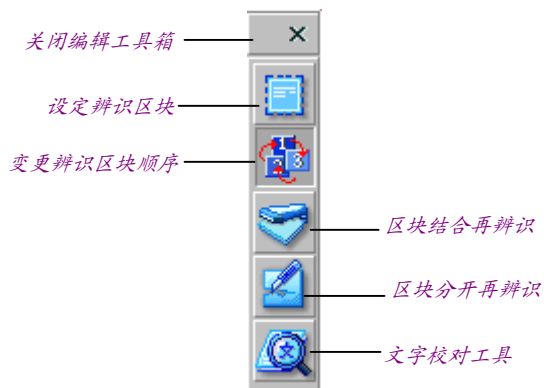
依您所点选的文字而移动。

快速浏览

选择此项，窗口内的候选字将依光标移动而变化，可供您快速浏览；若您没有选择此项，窗口则会在您点选某文字时才出现该字的候选字。

编辑工具箱

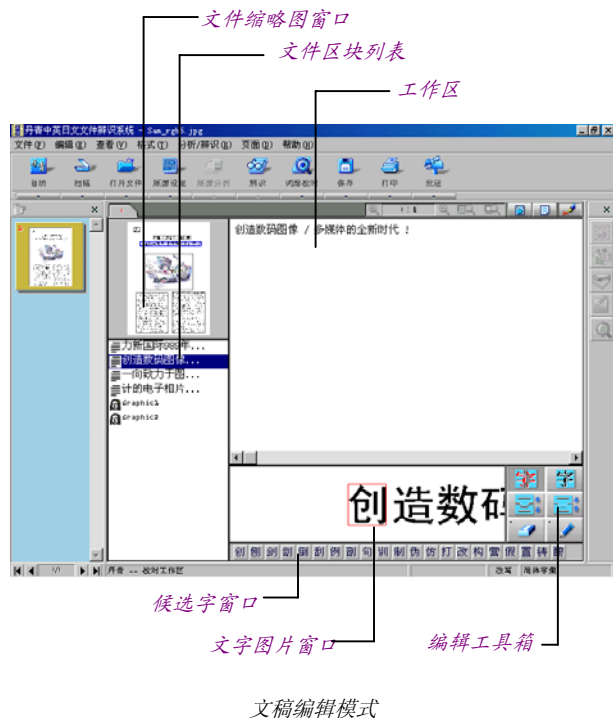
编辑工具箱提供您在全页图文模式中，所需的校对编辑工具。



全页图文模式中的编辑工具箱

文稿编辑模式：

在“文稿编辑模式”中最重要的工作就是核对系统辨识后的文稿。“文稿编辑模式”可供您查看辨识后的文字，并可针对文字进行校正的工作。丹青系统提供多种校对文稿的功能，如候选字、分/合字再辨识、分/合行再辨识、校对词库等，您也可以直接输入正确的字来进行校对。

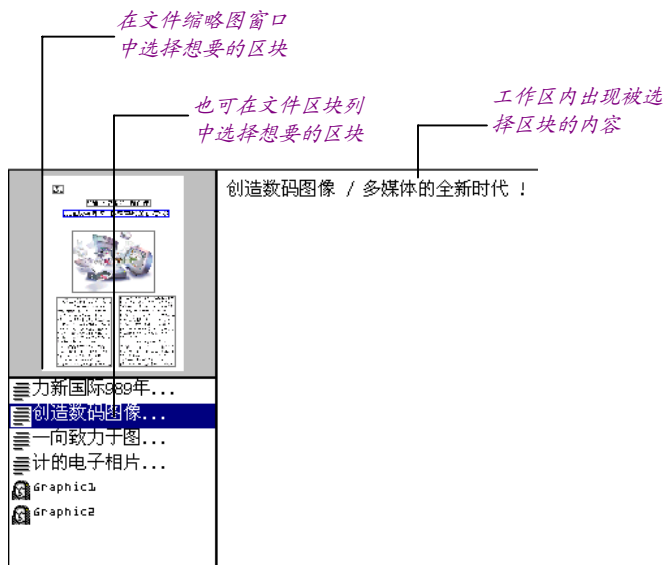


工作区

工作区内显示辨识后的文字；蓝色字样代表的是系统在辨识时遇到的疑问字。

文件缩略图窗口和文件区块列表

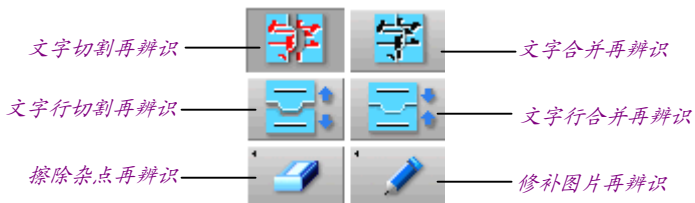
当您光标移至文件缩略图窗口时，您所点选的区块将会显示在工作区中。您也可以在文件区块列表中选择想要的区块；该区块的内容将会直接显示于工作区中。



文件缩略图窗口和文件区块列表

编辑工具箱

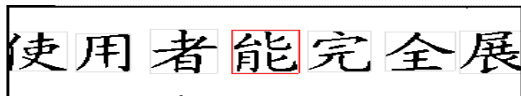
编辑工具箱提供您在文稿编辑模式中，所需的校对编辑工具。



文稿编辑模式中的编辑工具箱

文字图片窗口

文字图片窗口能放大显示光标所指的文字的原图图片，方便您校对与编辑文稿。



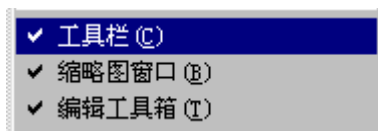
候选字窗口

您可在候选字窗口中点选正确的文字，以替换系统辨识时所误认的字。



改变屏幕配置

部分屏幕上的窗口可依需要而显示或隐藏，您可以选择“查看”菜单下的“工具栏”、“缩略图窗口”、或“编辑工具箱”命令来改变屏幕上的显示内容。



缩放图片显示比例

窗口内的图片可被放大或缩小至您所需要的显示比例，使您在编辑时更加得心应手。您可以选择“查看”菜单下的“与窗口同宽”、“全页显示”、或“实际大小”命令，或是“缩小显示”、“放大显示”、“缩小”、“放大”等命令来改变图片显示的大小。

页面移动及信息查询

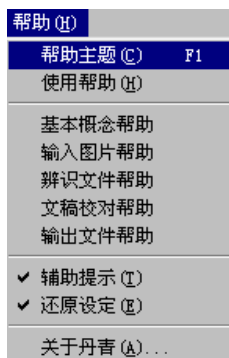
当您同时编辑数页文稿时，您可以选择“文件”菜单下的命令迅速地跳至特定的页面，并可选择“本页信息”命令来查询相关图片的大小及解析度。



“图片信息”对话框

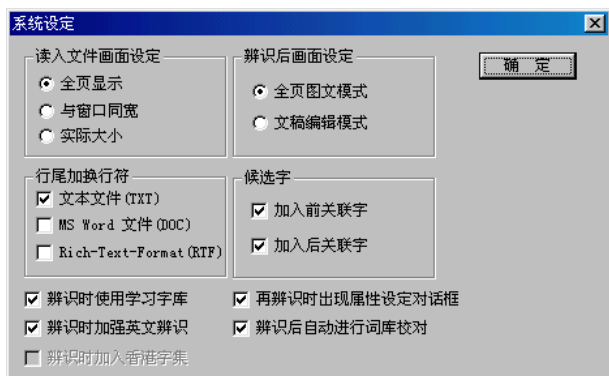
使用在线帮助

若您需要快速获得相关命令或功能的帮助与介绍，请按“帮助”菜单下的命令，即可获得您所需的信息。



设定系统默认值

您可选择“文件>系统预设”命令来设定丹青系统的默认值，“系统设定”对话框如下：



“系统设定”对话框

读入文件画面设定

选择最适合您的画面显示比例，如“全页显示”、“与窗口同宽”、“实际大小”等

辨识后画面设定

选择辨识后屏幕切换至“全页图文模式”或“文稿编辑模式”

行尾加换行符

选择辨识后需要在每行行尾加上换行符的文件格式

候选字

选择候选字的依据标准为前关联字或后关联字

前关联字:以被选字之前的字为依据，在词意上相关的字。以“中英文”一词为例，若英为被选字，则以“中”为依据的词意关联字有：中心、中华、中国、中央等

后关联字:以被选字之后的字为依据，在词意上相关的字。以“中英文”一词为例，若英为被选字，则以“文”为依据的词意关联字有：俄文、欧文、古文、公文等

辨识时使用学习字库

在辨识时使用学习字库，可提高辨识的正确率

辨识时加强英文辨识

在辨识中英文混合的文件时，加强英文的辨识

识别时加入香港字集	加入香港现行文件中特殊字的识别词汇
再辨识时出现属性设定对话框	在执行再辨识功能之前，出现属性设定对话框，可更改相关的设定
辨识时自动做词库校对	辨识时用内建的常用词库自动校对文稿

第3章

输入图片

所有要辨识的图片都必须先输入到丹青系统的图片处理窗口，才能做更进一步的辨识处理。因此，在辨识之前，您必须先知道如何将图片输入到丹青系统。

扫描与打开

大部分待辨识的图片都经由扫描过程而取得，您可在丹青系统内先设定扫描仪，并将扫描后的图片直接输入。您也可以直接打开磁盘内的图片进行辨识。

若要设定扫描仪：

1. 选择“文件>扫描仪设定”。
2. 在“扫描仪设定”对话框中选择您的扫描仪驱动程序之后，按“确定”。

若要输入图片：


丹青系统可识别以 TIF (G3, G4, PackBits), PCX, BMP 以及 JPG 文件格式所储存的影像。

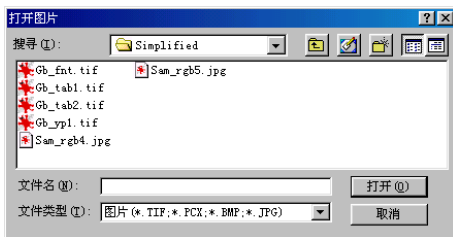
注意：若一个 TIF 档案内含有多页影像，丹青只读取第一页的影像。

您可自下列选择任何一种方式将文件输入：

- 选择工具栏上的“扫描”图标或“文件>扫描文件”命令。

系统会打开扫描界面，直接从扫描仪输入图片。（关于使用扫描界面，请参考扫描仪使用手册。）

- 选择工具栏上的“打开文件”图标  或“文件>打开图片”命令。



“打开图片”对话框


在“打开图片”对话框中选择一个图片，之后按“打开”，将选择的图片输入到丹青系统。

您也可以配合键盘上的“Shift”或“Ctrl”键，选择多份文件同时打开。

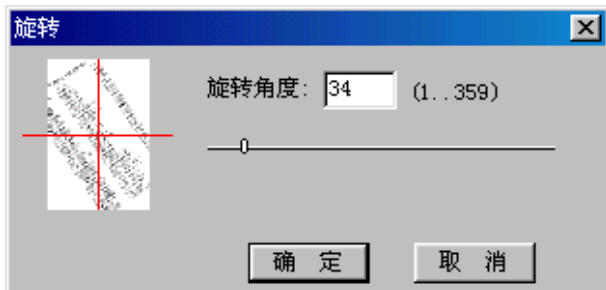
图片处理

一张品质良好的文件图片，是获得最佳辨识效果的关键因素。因此，在辨识前您应该查看输入的文件图片，并依情况做适当的图片处理，如转正倾斜的文件、除杂点、补漏白、切除不需辨识的部分、反白文件等，以提高辨识的正确率。

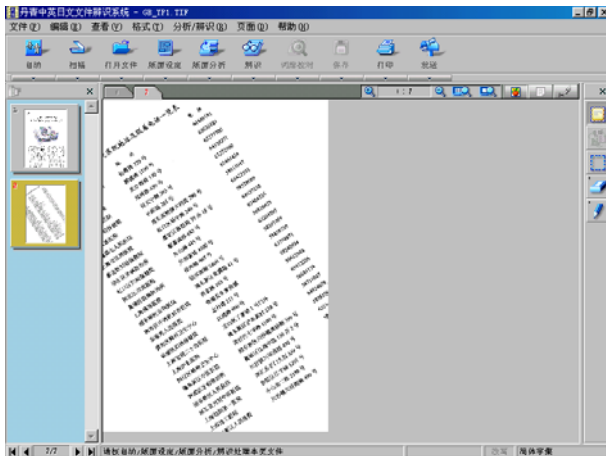
旋转图片

- 若图片倾斜角度小于 3 度，此为正常辨识可接受的范围，您不需调整图片角度。
- 若图片倾斜角度为 90 度，可利用“编辑>旋转>顺(逆)时针旋转 90 度”的命令，将文件图片转正。
- 若图片倾斜角度为 180 度，可选择“编辑>旋转 180 度”，将文件图片转正。
- 若您要一次转正多页文稿，可按一下缩略图窗口上方的图标，在出现的菜单中，选择您要的命令，系统将会一次转正目前在丹青系统中所打开的全部图片。

- 若图片倾斜角度大于 3 度且小于 10 度，可利用“编辑>任意角度旋转”功能，系统会出现如下对话框，自动为您侦测文件图片应转正的角度。





- 若图片的倾斜角度大于 10 度以上（如下图），建议您重新扫描图片。




倾斜角度过大的文件图片

清除杂点及补漏白


若文件图片上有杂点（尤其是在文字区域附近，与文字大小相近的杂点），可利用编辑工具箱上的“橡皮擦”工具将之去除；若图片上有漏白的部分，也可以使用“绘笔”工具补上，以提高正确率。

切除

若输入的图片不需全部辨识，您可利用编辑工具箱上的“选择图片区块”工具框选欲保留的区块，再选择“编辑>切除”命令，即可将不需要的部分切除。

反白功能

由于丹青系统无法辨识黑底白字的图片，若您输入的正是此类图片，可利用“编辑>反白”功能，将图片转换成白底黑字之后，再进行辨识。

若您要反白多页文稿，可按一下缩略图窗口上方的图标，在出现的菜单中，选择“全部反白”命令，系统将会一次反白目前在丹青系统中所打开的全部图片。

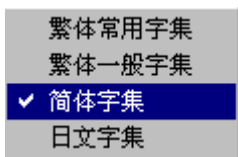
第4章

辨识文件

在执行文字辨识之前，您可以先做好一些辨识前的准备工作，如选择辨识字集、框选辨识区块、设定版面格式、执行版面分析、指定校对词库等，使丹青系统在辨识时更快速而准确。此外，丹青也提供自动辨识文件的功能，从输入以至辨识等各项流程皆能自动执行，让您轻松地获得想要的辨识结果。

设定辨识字集

当欲辨识文件里包含中文文字时，您可以选择“格式>设定辨识字集”，指定适合的辨识字集作为丹青系统辨识时的依据。




繁体常用字集	适合辨识一般白话文文件，如报纸、杂志等。
繁体一般字集	适合辨识包含文言文的文件，如经文、古书、典籍等。选择此项后的辨识速度会较使用“繁体常用字集”慢。
简体字集	适合辨识简体中文文件。
日文字集	适合识别日文文件。

设定辨识区块

若您要辨识整份文件，在执行辨识之前并不需要设定辨识区块。若您只想辨识部分文件，则可先设定该部分为辨识区块，点选该区之后再执行版面分析、辨识等工作，系统将只辨识您所框选的部分。

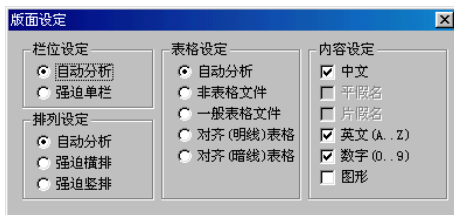
设定辨识区块的步骤如下：

1. 选择编辑工具箱上的“设定辨识区块”工具.
2. 拖动鼠标，框选欲辨识的图片区块即可。

若您要辨识数个辨识区块内的文字，可先分别设定各个欲辨识区块，之后再执行辨识，系统将会辨识所有设定的辨识区块。

设定版面格式

版面格式的设定主要在于设定欲辨识文件的属性，包括文件的横/竖排、单/多栏、所使用的语言及表格相关的设定等等。选择“格式>版面设定”命令之后，请依文件的内容选择所需要的设定。



“版面设定”对话框

栏位设定

栏位设定的默认值为“自动分析”，也就是让系统自动侦测图片上文字部份的栏位格式。在下列情况下可强迫设定为单栏：

- 排版较稀疏的单栏文件（如列表式文件），请在文件辨识之前，选择“强迫单栏”辨识图片上的文字。

- 图片上的文字部份为多栏位，但想存成单栏格式的文本文件（如通讯录），请在文件辨识之前，选择“强迫单栏”辨识图片上的文字。

排列设定

排列设定的默认值为“自动分析”，也就是让系统自动侦测图片的文字排列方式。若您的文字排列方式较为特殊，可依需要选择“强迫竖排”或“强迫横排”，以获得正确的版面分析结果。若您选择“强迫竖排”，那么系统将以竖排文字的顺序辨识图片上的文字；若您选择“强迫横排”，那么系统则以横排文字的顺序辨识文字。也就是说，辨识结果会因为您选择不同的文字排列方式而不相同。

表格设定

表格设定的默认值为“自动分析”，也就是让系统自动侦测图片为非表格文件或文件及其所属类型。若您的文件不含表格，您可自己设定为“非表格文件”；若您的文件包含表格，则可依表格类型选择“一般表格文件”、“对齐(明线)表格”或“对齐(暗线)表格”。

内容设定

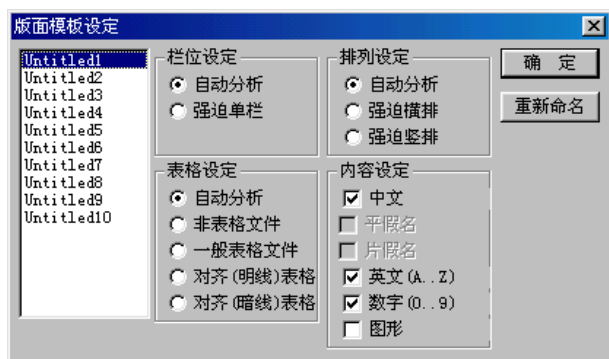
可选择图片所包含的内容属性，如图片为中/英/中英混合文件、以及是否含有数字或图形等。

设定版面模板

您可将经常需要设定的版面格式设定成版面模板，之后便可直接套用于相同类型的图片中，而不需一一重新设定。丹青系统保留十组版面模板，可供您直接使用，您也可以更改其中的设定资料。

若您要设定版面模板：


1. 选择“格式>版面模板设定”。
2. 在出现的对话框中，选择欲设定的模板名称，并设定相关的版面格式属性，之后按“确定”即可。



“版面模板设定”对话框

您也可以按一下“重新命名”，更改目前所选择的模板名称。

若您要套用版面模板：


1. 按一下工具栏的“版面设定”图标下方的下拉式菜单。
2. 在下拉式菜单中选择您想要的版面模板，即可将所选择的模板直接套用于目前打开的图片上。

版面分析


执行版面分析的目的在于将图形与文字图片区块分离，分割出待辨识的区块，并决定辨识区块的顺序，以利系统辨识。您可以让系统自动执行版面分析，或是自己设定区块及辨识顺序；除此之外，您也能在执行版面分析之后，分别设定各个区块的属性并将版面保存起来，当需要辨识相同版面的文件时，便可直接套用。

自动分析版面



按一下功图标列上的“版面分析”，系统便会自动分离图文并分割所有待辨识区块内的图片文字。在执行版面分析后，原图片上会出现系统所分析出的区块框线。


手动设定版面

1. 选择编辑工具箱的“设定辨识区块”工具.
2. 拖动鼠标，框选您要设成区块的部分。被框出的部分会以黄色显示。
3. 选择“格式>设成区块”，即可将所框选的部分设定成欲辨识的区块。

改变辨识顺序

在版面分析后所分割出的区块都有其被辨识时的顺序编号。若您选择自动分析版面，系统会依照设定的文字横/竖排方式来决定辨识区块的顺序；若您自己设定区块，辨识顺序则依每个区块设定的先后而定。当您按一下编辑工具箱的“变更辨识区块顺序”工具，每一个区块的左上角会出现辨识时的顺序编号。辨识顺序的设定将会影响系统辨识后的文字内容排列方式。

变更辨识区块顺序的步骤如下：

1. 在编辑工具箱上选择“变更辨识区块顺序”工具.
2. 在欲改变辨识顺序的区块上按住鼠标，并拖动鼠标指标到另一个想改变成其辨识顺序编号的区块上；例如想改变区块 3 成为区块 2，只要将鼠标自编号 3 号的区块拖动到编号 2 号的区块。您会发现，当您放开鼠标指标的同时，辨识区块左上角的排列顺序也跟着改变了。

您可以根据您的需要，设定各个区块的辨识顺序。

保存版面

您可保存经常使用的版面，于下次辨识时直接套用，可节省系统执行版面分析的时间。若您要保存版面，选择“格式>保存版面”；若您要使用既有的版面，选择“格式>打开版面”，或是直接在“版面分析”图标的下拉式菜单中选择。


辨识文件

在完成输入图片、设定辨识字体、设定辨识区块及变更辨识区块顺序等步骤之后，系统便可以根据您的设定开始辨识文件。

当完成辨识工作后，系统会自动进入“全页图文模式”或“文稿校对窗口”，让您校对辨识后的文本文件。您可以选择“文件>系统设定”，在“系统设定”对话框中指定辨识后所出现的画面显示模式。

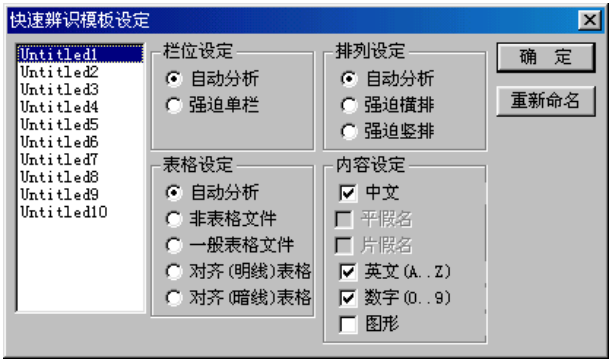
若您经常辨识某类文件，设定快速辨识模板将为您省却一一设定的工作，并能快速地呈现辨识后的结果。

若您要辨识文件：

按一下工具栏上的“辨识”图标 。

若您要设定快速辨识模板：


1. 选择“分析/辨识>快速辨识模板设定”。
2. 在出现的对话框中，选择欲设定的模板名称，并设定相关的版面格式属性，之后按“确定”即可。



“快速辨识模板设定”对话框

您也可以按一下“重新命名”，更改目前所选择的模板名称。您所设定的模板将会出现于“辨识”图标下方的下拉式菜单中。

若您要套用快速辨识模板：

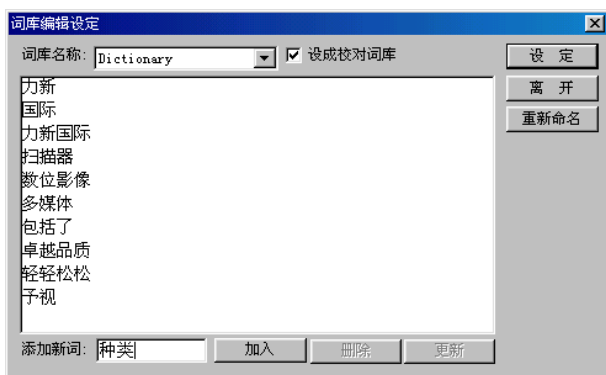
在“辨识”图标  的下拉式菜单中，选择想要的快速辨识模板即可。

选择校对词库

校对词库里包含您常用的词汇；在辨识的过程中系统将依您所选择的校对词库执行辨识。因此，依需要设定不同种类的词库，在辨识不同种类的文件时，将更节省您在辨识及校对时所花费的时间。

设定校对词库

1. 选择“分析/辨识>词库设定”。
2. 依您的需要设定下列选项：




词库设定对话框

重新命名	更换所选择词库的名称
插入	将您输入于“添加新词”文字框内的符号或字词加至个人词库中
删除	将您在词库中所选择的字词删除
更新	更改您在词库中所选择的字词
设定校对词库	将所选择的词库设定成系统校对时使用的词库，您并可同时选择多种词库为校对词库

使用词库校对


您可更换不同的词库，重新执行校对，使辨识结果更令人满意。使用词库校对的步骤如下：

1. 选择“分析/辨识>词库设定”，指定您要的校对词库，按“设定”即可。
2. 按一下工具栏上的“词库校对”图标 ，系统将依据您所指定的词库再次辨识。

自动辨识文件

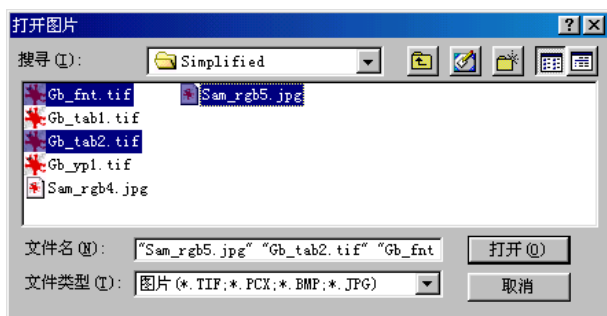
丹青提供自动辨识文件的功能，您可以随着自动引导模式一一设定从输入以至辨识等各项流程所必要的选项，之后由丹青为您执行自动辨识的工作。此外，您也能设定自动辨识模板，同时执行多份文件的自动辨识工作。

自动辨识

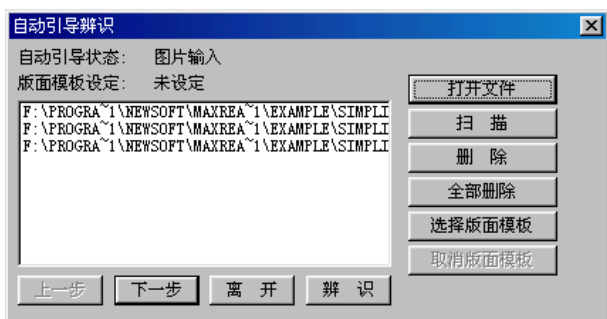
1. 按一下工具栏上的“自动”图标 。
2. 在出现的“自动引导辨识”对话框中，选择“打开文件”。



3. 选择您要打开的图片。您可同时选择打开多份文件，这些文件的名称会列在对话框中。若您不需要其中的部分文件，也可选择“删除”或“全部删除”。



4. 选择“下一步”，依次设定各个选项。



5. 重复第四步骤的动作。当您完成所有的设定时，系统将开始执行辨识。

设定自动模板

您可事先设定自动辨识的属性，使系统的自动辨识结果更符合您的需要。当您使用自动辨识功能时，系统将依据您所设定的自动模板执行辨识。

设定自动模板的步骤如下：

1. 选择“文件>自动模板设定”。
2. 在“自动模板设定”对话框中，设定下列选项，之后按“确定”即可。



自动模板设定对话框

重新命名	更换所选择模板的名称
新增页	增加新的原稿图片于模板中
删除页	删除模板中所选择的原稿图片
全部删除	删除模板中全部的原稿图片
栏位设定	设定模板或某个被选择的图片的栏位属性
排列设定	设定模板或某个被选择的图片的文字横竖排
表格设定	设定模板或某个被选择的图片的表格性质
内容设定	设定模板或某个被选择的图片所包含的资料内容
选择版面模板	单击“浏览”，可选择合适的版面文件(*.tpl)，并应用于模板中的所有文件

第5章

文稿校对

当系统完成辨识后，您可在“全页图文模式”及“文稿编辑模式”中查看辨识结果。系统可显示在辨识时遇到的疑问字，疑问字也许是原图片文字的模糊不清或是版面设定、分析有误所导致，您可以根据实际情况校对文稿。以下则分别说明如何校对辨识后的文件。


放弃辨识

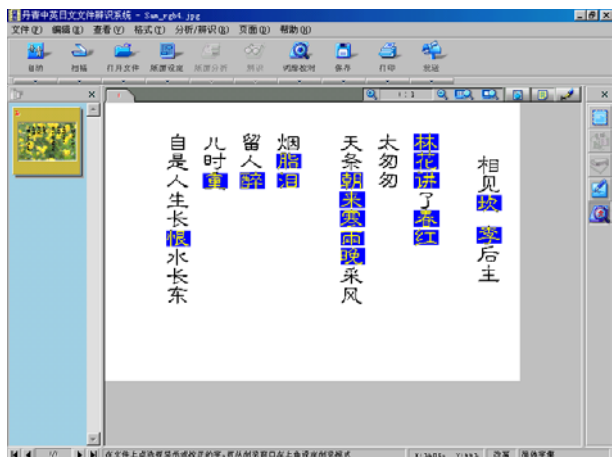
若您在辨识后想放弃辨识结果并重新设定，请选择“分析/辨识>放弃本页辨识”或“分析/辨识>放弃全部辨识”将辨识结果消除。

校对文稿

当辨识完成后，画面会依您的系统默认值出现“全页图文模式”或“文稿编辑模式”。“全页图文模式”可供您观看辨识后的文稿版面全貌，而“文稿编辑模式”则能分段显示辨识后的结果。您可以在查看工具箱上选择这两种不同的显示模式。

若您要在“全页图文模式”中校对文稿：

1. 按一下编辑工具箱上的“文字校正”工具，文稿中的疑问字会以蓝底黄字的字样显示。



蓝底黄字的字样显示出疑问字

2. 使用“文字校正”工具在第一个疑问字上点一下，并在出现的“文字校正”窗口中选择正确的字。您所选择的字将会替换指定的疑问字。



“文字校正”窗口

若在“文字校正窗口”中找不到您要的替代字，您也可以使用一般的键盘输入法将文字输入。

3. 同时按下键盘上的“Shift”及“F3”键，将光标移到下一个疑问字元，或使用“文字校正”工具选择任何辨识错误的字，继续进行文字的校对即可。

若您要在“文稿编辑模式”中校对文稿：

1. 在文件浏览图中直接点选欲校对的区块，或是在文件区块列表中指定区块，该段的辨识结果将出现于工作窗口内，且文稿中的疑问字会以蓝色字样显示。
2. 用鼠标点一下工作窗口内系统辨识错误的字，在“文字图片窗口”中，也会同时会用红线框出其对应的原字元图片。
3. 您可以在“文字图片窗口”下方的“候选字窗口”中，点选一个正确的候选字来替代辨识错误的字。您所选择的候选字会替换掉工作区内鼠标光标所在位置后的字元。

若在“候选字窗口”中找不到您要的替代字，请将光标移到工作区中辨识错的文字左方，使用键盘输入法直接输入您需要的字。


4. 同时按下键盘上的“Shift”及“F3”键，将光标移到下一个疑问字元，或移动鼠标选择任何辨识错误的字，继续进行文字的校对即可。

再辨识

某些图片可能无法使系统做出正确的分割，并因而造成辨识上的错误。您可以使用擦除杂点、分/合字再辨识、分/合行再辨识与分/合区块再辨识的功能，重新辨识错误结果。


擦除杂点再辨识

去除图片上的杂点可以提高系统辨识的正确率。当系统已进入文稿校对窗口后，您可以使用“擦除杂点再辨识”工具擦除文字图片窗口中红框内的字元。擦除的步骤如下：

1. 在工作区中选定系统辨识错误的字，文字图片窗口中也会同时会出现以红框框住的对应字元。
2. 选择编辑工具箱中的“擦除杂点再辨识”工具.
3. 拖动光标，将图片窗口内的杂点擦除干净。之后在红框外按一下或按“Enter”键，系统会重新辨识该图片字元。您也可以按“Esc”键复原之前所擦除掉的部分。


修补图片再辨识

增补图片上的漏白部分可以提高系统辨识的正确率。当系统已进入文稿校对窗口后，您可以使用“修补图片再辨识”工具增补文字图片窗口中红框内的字元。补漏白的步骤如下：

1. 在工作区中选定系统辨识错误的字，文字图片窗口中也同时会出现以红框框住的对应字元。
2. 选择编辑工具箱中的“修补图片再辨识”工具。
3. 拖动鼠标指标，将图片窗口内的漏白部分补上。之后在红框外按一下或按“Enter”键，系统会重新辨识该图片字元。您也可以按“Esc”键复原之前补漏白的部分。


文字切割再辨识

将相邻二个或数个辨识错的字元分开并予以重新辨识。使用“文字切割再辨识”工具的步骤如下：

1. 用鼠标点一下工作区内辨识错误的字元，文字图片窗口中也同时会出现用红线框出的对应字元图片。
2. 选择编辑工具箱中的“文字切割再辨识”工具。
3. 在红框中按住鼠标左键，图片字元会被分割成二个部份。当您调整好红线切割的位置时放开鼠标，系统便会重新辨识被切割的两个字元，同时，工作区内的辨识文字也会跟着更新。

文字合并再辨识


将相邻二个或数个辨识错的字元合并并予以重新辨识。使用“文字合并再辨识”工具的步骤如下：

1. 用鼠标点一下工作区内辨识错误的字元，文字图片窗口中也同时会出现用红线框出的对应字元图片。
2. 选择编辑工具箱中的“文字合并再辨识”工具。

3. 在红框中按住鼠标左键，拖动至欲合并的字后放开鼠标按键，系统会重新辨识合并部份的字元，在此同时，工作区内的辨识文字也会跟着更新。


文字行切割再辨识

将因二行相连而辨识错的文字分开予以重新辨识。使用“文字行切割再辨识”工具的步骤如下：

1. 用鼠标点一下工作区内辨识错误的文字，文字图片窗口中也同时会出现用红线框出的对应部分。
2. 选择编辑工具箱中的“文字行切割再辨识”工具.
3. 在红框中按一下鼠标左键，相连文字行会被分割开并重新辨识，而工作区内的辨识文字也会跟着更新。



文字行合并再辨识

将被错误分割成二行的文字合并予以重新辨识。使用“文字行合并再辨识”工具的步骤如下：

1. 用鼠标点一下工作区内辨识错误的文字，文字图片窗口中也同时会出现用红线框出的对应部分。
2. 选择编辑工具箱中的“文字行合并再辨识”工具.
3. 在红框中按住鼠标左键，拖动到欲合并的文字行后放开鼠标按键，系统会重新辨识合并部份的文字行，在此同时，工作区内的辨识文字也会跟着更新。

区块再辨识



当您发现某区块的版面分析错误，如文字的横竖排列错误或是中英语言设定错误时，您可以针对该区块再次辨识。区块再辨识的步骤如下：

1. 按一下查看工具箱上的“全页图文模式”工具.
2. 选择编辑工具中的“选择辨识区块”工具，并在欲重新辨识的区块上按一下鼠标左键。被选择的区块会以黄色显示。

3. 选择“格式>文件格式设定”，重新设定该区块的属性。
4. 选择“辨识/分析>区块再辨识”，系统将重新辨识指定的区块，并更新前次的辨识结果。



区块结合再辨识

可合并被错误分割的区块，再次辨识。

1. 按一下查看工具箱上的“全页图文模式”工具 。
2. 选择编辑工具中的“区块结合再辨识”工具 。
3. 按住鼠标左键并拖动鼠标，将欲合并的区块框住。
4. 放开鼠标，系统将重新辨识合并的区块，并更新前次的辨识结果。

区块分开再辨识

可分割被错误合并的区块，再次辨识。

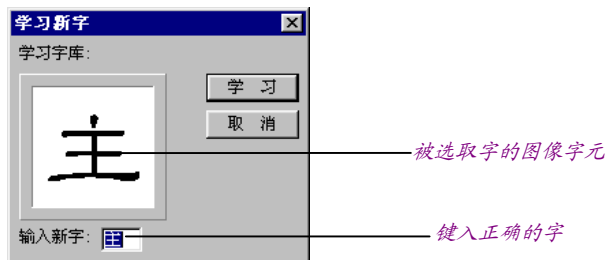
1. 按一下查看工具箱上的“全页图文模式”工具 。
2. 选择编辑工具中的“区块分开再辨识”工具  工具。
3. 按一下鼠标左键并调整产生的红线的位置(您也可拖动鼠标直接拉出红线)，将合并的区块分开。
4. 放开鼠标，系统将重新辨识被分割的区块，并更新前次的辨识结果。

学习新字

当在校对文稿时，若系统经常辨错某些文字，您可以使用“学习新字”的功能，将常辨识错的字元输入到学习资料库中，留待以后辨识时使用。您也可依文件的性质，设定各种不同的学习字库。

使用“学习新字”：

1. 选择“分析/辨识>学习新字”命令，屏幕上会出现一个“新字学习”的对话框。



在对话框的上方为被选取字的图片字元(也就是在“文字图片窗口”中被红框框选者)。

2. 在“新字”文字框中输入正确的字。
3. 按下“学习”键，将新字输入到学习字库中，并置换工作区内的错误字元。

删除学习字

若您想删除学习字库中的字，请选择“分析/辨识>删除学习字”命令。在学习字库中选择欲删除的字之后，按“删除”即可。

学习字设定

“学习字设定”功能可让您选择特定的学习字库做为系统的默认值，也可更改学习字库的名称。

1. 选择“分析/辨识>学习字设定”命令。



2. 指定您想要的预设学习字库，按“确定”。您也可选择“重新命名”，更改学习字库的名称。

第6章

输出文件

在丹青系统中您可保存辨识前的图片、常用的版面格式、以及辨识后的图文和表格，再加以使用及编辑。

保存辨识前的图片

若您想保留这些在辨识前经扫描仪或其他方式输入的图片，可将之保存成 BMP、TIFF、PCX、JPEG 等图片格式，方便以后再辨识利用。

保存图片：

1. 选择“文件>保存本页>保存本页原稿图片”命令。
2. 在“保存原稿图片”对话框中指定路径、文件名及文件格式，之后按“保存”即可。



“保存原稿图片”对话框

保存辨识后的图文和表格

丹青系统提供多种文件格式，可保存辨识后的图文及表格。您可依需要选择保存本页或保存整份文件，并将之保存成 TXT、DOC、RTF、XLS、SLK、CSV 等文件格式，在写字板、Word、Excel 等文字处理器中编辑。此外，您还可以将文件存成 HTML 格式，通过网络浏览器（如 Internet Explorer、Netscape Navigator 等）直接打开。

若您要保存辨识后的结果：

1. 若您要保存本页，选择“文件>保存本页>保存本页辨识结果”；若您要保存整份文件，选择“文件>保存文件辨识结果”。屏幕上将出现“保存文件辨识结果”对话框。



2. 指定路径、输入文件名，并选择您要的存档类型。

文件格式	说明
TXT	纯文本文件。选择此种文件格式，系统将只保存文件中的文字部分，而不会保留其中的图形。
DOC	Winword 文档。选择此种文件格式，系统会将文件保存成 doc 文档；若文件中包含图形，系统会将图形保存成 JPEG 图档并依序编号。当保存多页文件时，系统也会自动在各页面之间插入分页符号，以利区别。
RTF	可保留文字大小、版面位置及表格格线等信息的文件格式。若文件中包含图形，系统会将图形保存成 JPEG 图档并依序编号。当保存多页文件时，系统也会

	自动在各页面之间插入分页符号，以利区别。
XLS	可在 Excel 中打开的文件。
SLK	可在 Excel 中打开的文件。
CSV	可在 Excel 中打开的文件。
HTML	可通过网络浏览器打开的文件。若文件中包含图形，系统会将图形保存成 JPEG 图档并依序编号。当保存多页文件时，系统也会自动在各页面之间插入分页符号，以利区别。

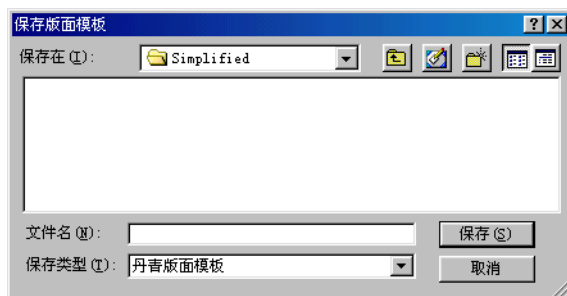
3. 按下“保存”键即可。

保存常用的版面格式

若您经常需要使用某一种版面格式，可将此版面格式保存成版面文件 (*.TPL)，应用于辨识前的文件图片，节省您再次设定的时间并提高辨识的正确率。

保存版面格式：

选择“格式>保存版面”命令，屏幕上将出现“保存版面”对话框。输入文件名后，按“保存”即可。



“保存版面”对话框

若您要套用某个保存的版面格式：

请选择“格式>打开版面”命令，在对话框中选择要套用的版面格式，按下“打开文件”按钮即可。


打印

丹青提供打印原稿及辨识结果的功能。若您需要打印辨识结果，可选择“文件>打印>打印辨识结果”；若需要打印原稿图文，可选择“文件>打印>打印原稿图片”。

发送

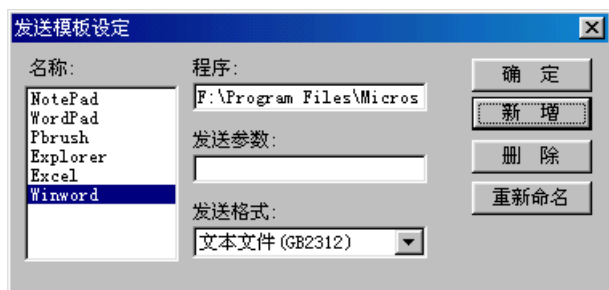
丹青系统提供直接发送功能，可将文件图片或辨识结果发送至电子邮件软件或其他相关的应用软件，如写字板、画图等。此外，丹青保留 10 组发送模板可让您事先设定，在其中指定发送的文件格式及欲打开的应用程序，之后便可直接套用模板，将文件发送至您想要的应用软件中。

若您要将文件发送成电子邮件：

按一下工具栏上的“发送”图标 ，便可将目前打开的文件发送至电子邮件软件中。

注意：您所安装的电子邮件软件必须支援 **MAPI** 的电子邮件系统，如 **Exchange**、**Outlook Express** 等，才能在丹青系统中直接发送电子邮件。



1. 选择“文件>发送模板设定”。
2. 在“发送模板设定”对话框中设定所需的选项：



“发送模板设定”对话框

新加	插入新的发送模板
删除	删除原有的发送模板
重新命名	更换原有发送模板的名称
程序	输入该应用程序的执行文件名
发送参数	输入该应用程序所需的参数
发送格式	指定传送出的文件格式

若您要套用发送的模板：

按一下工具栏“发送”图标下方的下拉式菜单，选择您要套用的发送模板即可。


注意：在发送的过程中，丹青系统会将文件暂存于 *Pccrtemp* 的目录里，您可以定期删除该目录内的所有文件，以节省您的电脑磁盘空间。

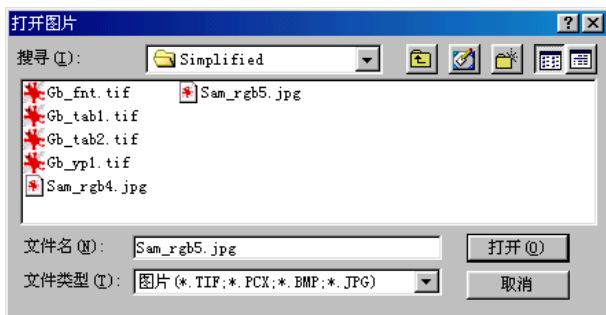
第7章

图文辨识范例

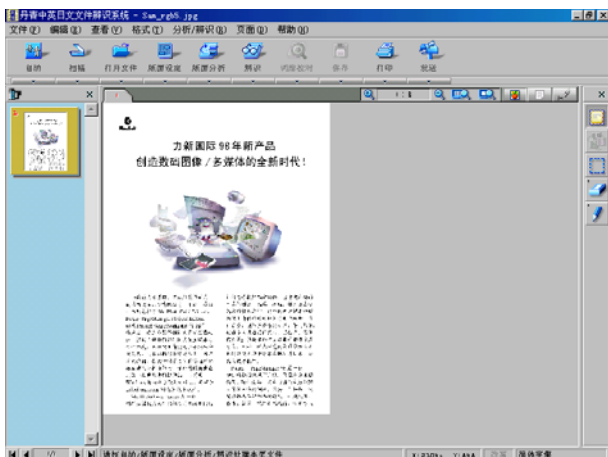
含有图形及文字的文件是一般最常见的文件类型，本章将为您示范图文文件的设定，辨识流程及其应用。

辨识含有图形及文字的文件

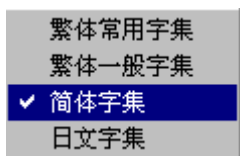
1. 按一下工具栏上的“打开文件”图标  打开文件。
2. 在出现的对话框中，选择丹青程序文件夹中 Example 目录里的 Sam_rgb5. jpg 文件。




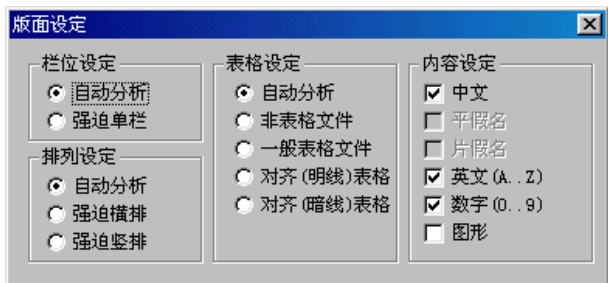
3. 按一下“打开”，将文件输入。






4. 选择“格式>设定辨识字集”，在二级菜单中选择“简体字集”或由状态栏的右方选择“简体字集”。

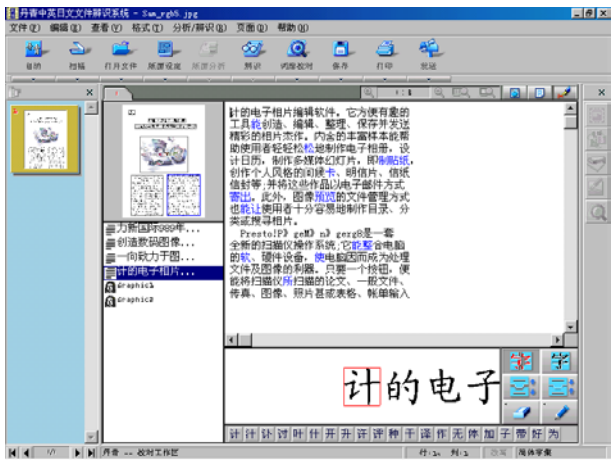




5. 按一下工具栏上的“版面设定”图标 。
6. 在“版面设定”对话框中，选择“栏位设定—自动分析”、“排列设定—自动分析”、“表格设定—自动分析”、“内容设定—中文、英文、数字”。



7. 按一下工具栏上的“辨识”图标 。系统将开始为您辨识文件。

8. 选择查看工具箱上的“全页图文模式”  及“文稿编辑模式”  工具查看辨识结果，并依需要校对文稿。（校对文稿的详细步骤请参考第 5 章）。



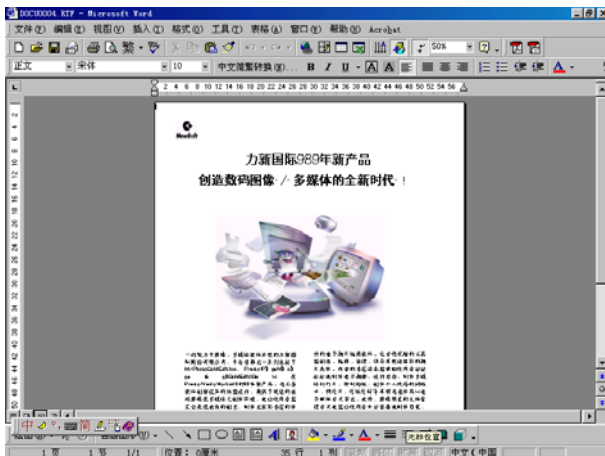
9. 在工具栏的“发送”图标  的下拉式菜单中 ，选择“Winword”。



注意：您必须先安装在电脑中 Winword 软件，才能直接发送辨识结果。并且在安装之后，要先将之设定为发送模板，才能在“发送”图标的下拉式菜单中直接选取套用。关于发送模板的设定，请参考第 6 章的说明。

注意：您可以选择“文件>系统设定”，在对话框中设定 Winword 文档加上换行符，便可保留文件原来的版面样式。若您想要再编辑该文件，则不要选择加上换行符，可方便文件的再编辑。

10. 系统将会直接发送辨识结果至 Winword 软件中。在 Word 中选择“查看>整页显示”，即可浏览辨识后的文件全貌。




发送至 Winword 中的辨识结果

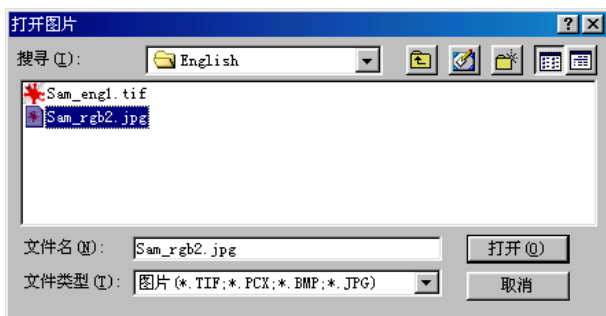
第8章

英文辨识范例

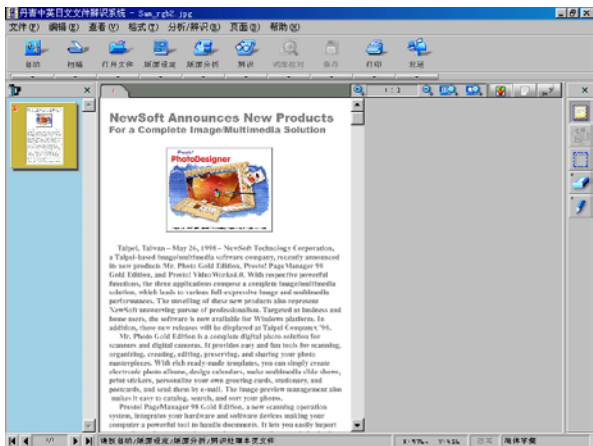
除了中文文件外，丹青也能为您辨识英文文件。以下将为您介绍英文文件的辨识步骤。


辨识英文文件

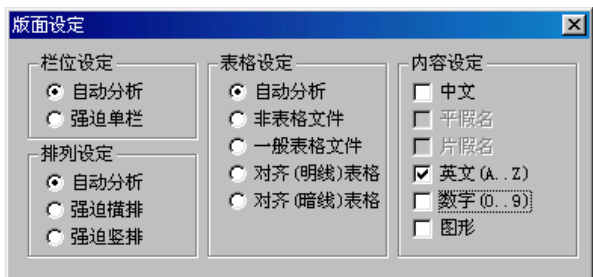
1. 按一下工具栏上的“打开文件”图标  打开文件。
2. 在出现的对话框中，选择丹青程序文件夹中 Example 目录里的 Sam_rgb2.jpg 文件。






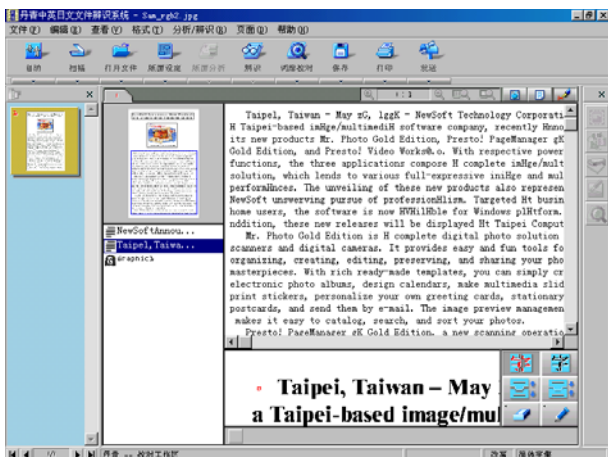
3. 按一下“打开”，将文件输入。




4. 按一下工具栏上的“版面设定”图标 。选择“栏位设定—自动分析”、“排列设定—自动分析”、“表格设定—自动分析”、“内容设定—英文”。



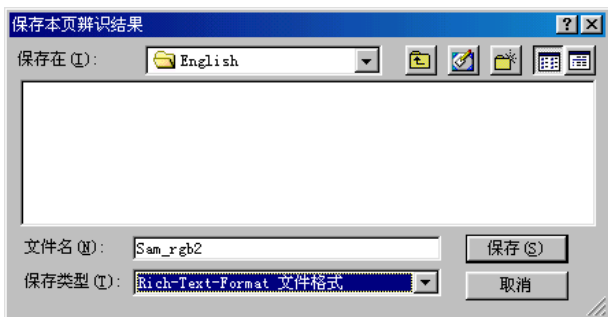
5. 按一下工具栏上的“辨识”图标 。系统将开始为您辨识文件。选择查看工具箱上的“全页图文模式”  及“文稿编辑模式”  工具查看辨识结果，并依需要校对文稿。(校对文稿的详细步骤请参考第 5 章)。




6. 校对完成后，您可在工具栏的“保存”图标  的下拉式菜单中，选择“保存本页辨识结果”。

保存文件辨识结果
保存本页辨识结果
保存本页原稿图片

7. 在保存对话框中，输入文件名称 Sam_rgb2，并指定存档类型为 RTF 文件格式。

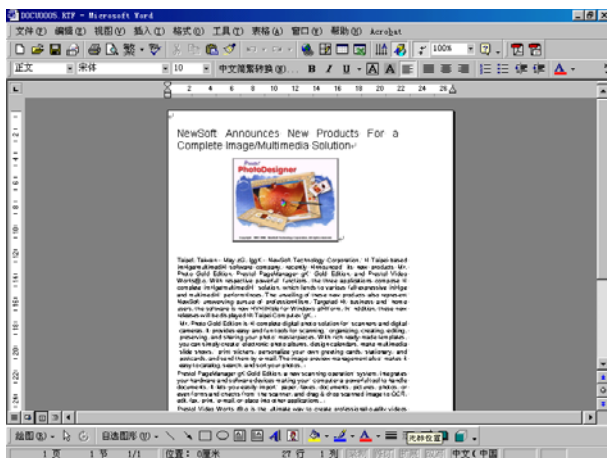


注意：您可选择“文件>系统设定”，在对话框中设定 RTF 文档加上换行符，便可保留文件原来的版面样式。若您想要再编辑排版该文件，则不要选择加上换行符。

8. 您也可以直接发送文件至相关的软件中。可在工具栏的“发送”图标的下拉式菜单中，选择“Winword”。

注意：您必须先在电脑中安装 Winword 软件，才能直接发送辨识结果。并且在安装之后，要先将之设定为发送模板，才能在“发送”图标的下拉式菜单中直接选取套用。关于发送模板的设定，请参考第 6 章的说明。

9. 系统将会直接发送辨识结果至 Winword 软件中。在 Winword 中选择“查看>整页显示”，即可浏览辨识后的文件全貌。



在 Winword 中的辨识结果


注意：若您不需要保存图形，也可选择 txt 纯文本文件格式来保存文件，所保存的文件将只包含文件中的文字内容。您可将该纯文本文件发送至写字板或记事本等文字软件中。

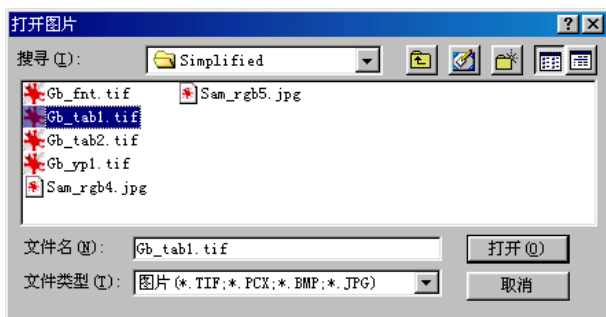
第9章

表格辨识范例

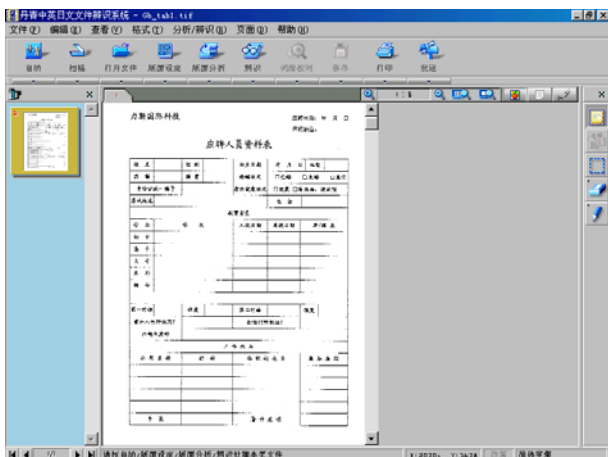
除了图文并存及中英文文件的辨识外，丹青也能为您辨识各式各样的表格，如公文、通讯录、履历表、成绩单等。表格的形式及文字皆能保持原貌；您也能将表格的辨识结果发送至 Word、Excel 等软件中再进一步编辑处理。

辨识一般表格图片

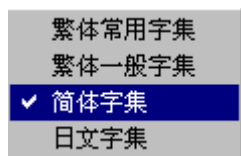
1. 按一下工具栏上的“打开文件”图标  打开文件。
2. 在出现的对话框中，选择丹青程序文件夹中 Example 目录里的 GB_TAB1.tif 文件。




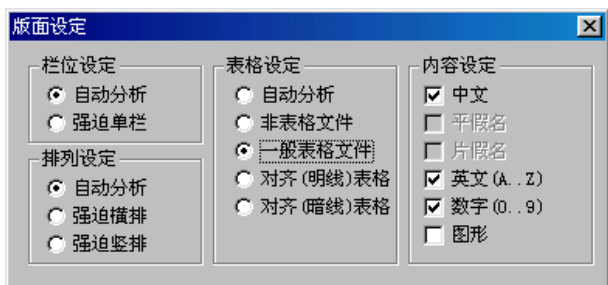
3. 按一下“打开”，将文件输入。






4. 选择“格式>设定辨识字集”，在二级菜单中选择“简体字集”或由状态栏的右方选择“简体字集”。

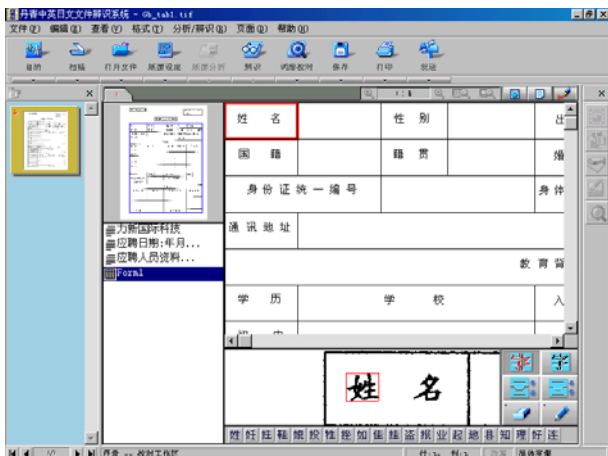



5. 按一下工具栏上的“版面设定”图标 。
6. 在“版面设定”对话框中，选择“栏位设定—自动分析”、“排列设定—自动分析”、“表格设定—一般表格文件”、“内容设定—中文、英文、数字”。

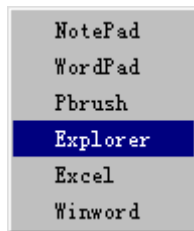


7. 按一下工具栏上的“辨识”图标  辨识。系统将开始为您辨识文件。

8. 选择查看工具箱上的“全页图文模式”  及“文稿编辑模式”  工具查看辨识结果，并依需要校对文稿。（校对文稿的详细步骤请参考第 5 章）。



9. 您可以直接发送文件至相关的软件中。可在工具栏的“发送”图标  的下拉式菜单中，选择“Explorer”。




注意：您必须先安装在电脑中安装网络浏览器，才能直接发送辨识结果。并且在安装之后，要先将之设定为发送模板，才能在“发送”图标的下拉式菜单中直接选取套用。关于发送模板的设定，请参考第 6 章的说明。

10. 系统将会直接发送辨识结果至 Internet Explorer 网络浏览器中。

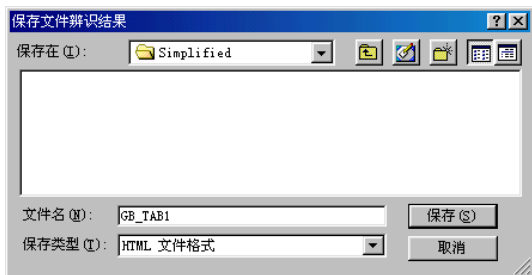


在 Internet Explorer 网络浏览器中的辨识结果

11. 您也可以选择保存文件。可在工具栏的“保存”图标  的下拉式菜单中，选择“保存本页辨识结果”。


保存文件辨识结果
保存本页辨识结果
保存本页原稿图片

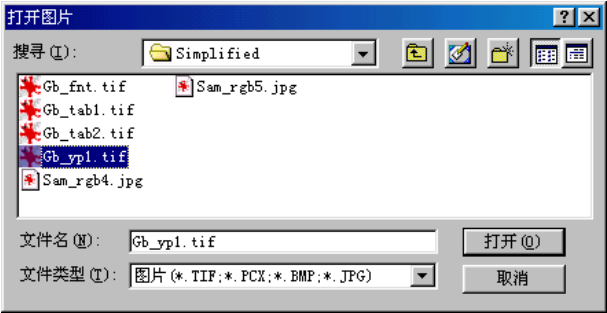
12. 在保存对话框中，输入文件名称 GB_TAB1，并指定存档类型为 HTML 文件格式。系统会将文件保存成 HTML 格式，并将文件中的图形依序编号保存成*. JPEG 文件。



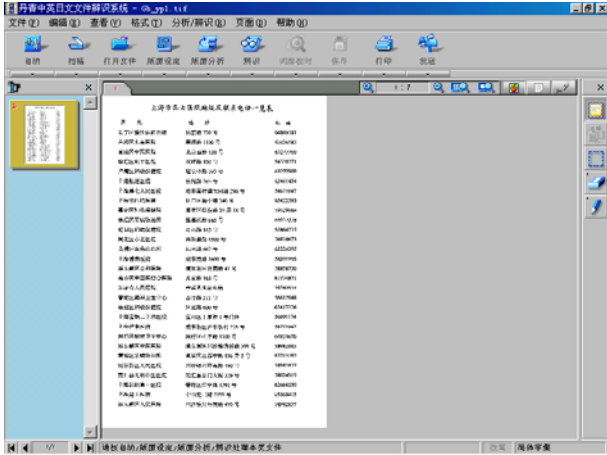
保存辨识结果

辨识暗线表格图片

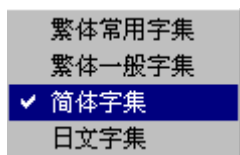
1. 按一下工具栏上的“打开文件”图标 。
2. 在出现的对话框中，选择丹青程序文件夹中 Example 目录里的 Gb_YPl.tif 文件。




3. 按一下“打开”，将文件输入。

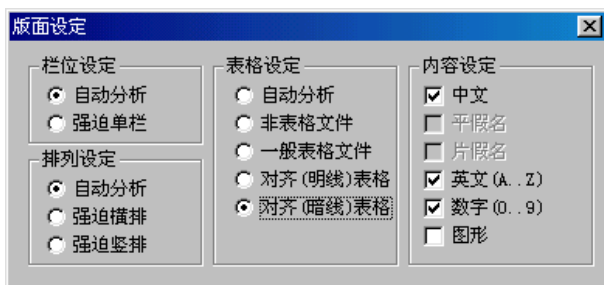



4. 选择“格式>设定辨识字集”，在二级菜单中选择“简体字集”或由状态栏的右方选择“简体字集”。

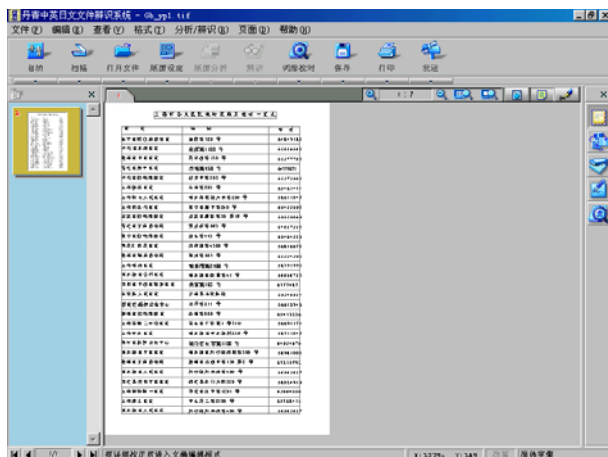




5. 按一下工具栏上的“版面设定”图标 。

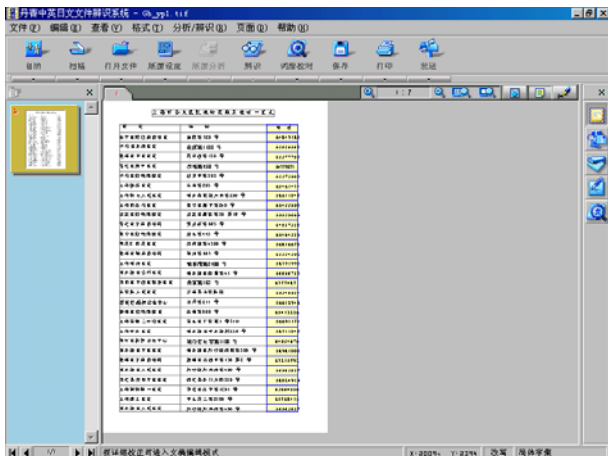
6. 在“版面设定”对话框中，选择“栏位设定自动分析”、“排列设定—自动分析”、“表格设定—对齐(暗线)表格”、“内容设定—中文、英文、数字”。



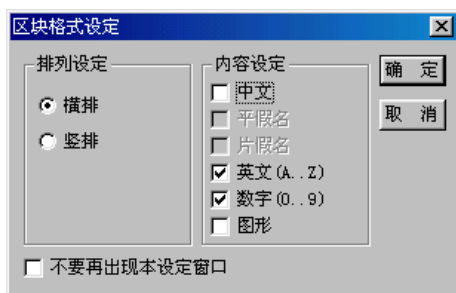
7. 按一下工具栏上的“辨识”图标 。系统将开始为您辨识文件。





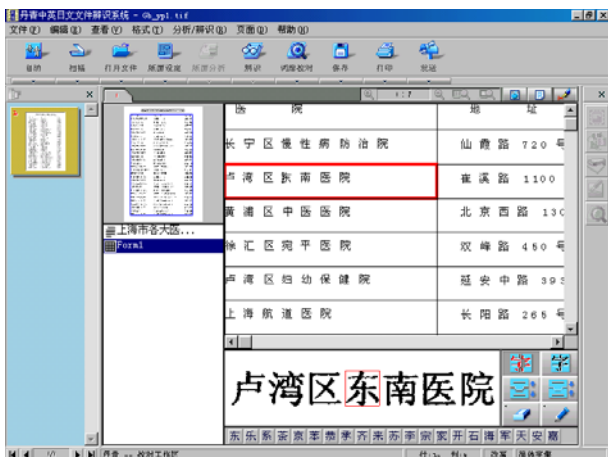
8. 选择查看工具箱上的“全页图文模式”及“文稿编辑模式”工具查看辨识结果，并依需要校对文稿。（校对文稿的详细步骤请参考第 5 章）。
9. 由于电话号码区块无中文字，所以这里重新设置区块的属性将会提高识别结果的准确性。请选择编辑工具箱上的“选择识别区块”工具，利用鼠标选择要重新识别的区块。被选择的区块会显示黄色。



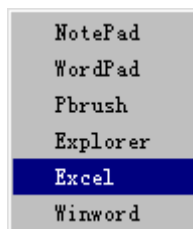
10. 选择“分析/辨识>区块再辨识”。在“内容设定”选项中，取消“中文”设置。点击“确定”，系统将开始为您重新识别此区域。



11. 选择查看工具箱上的“全页图文模式”及“文稿编辑模式”工具查看辨识结果，并依需要校对文稿。（校对文稿的详细步骤请参考第 5 章）。



12. 您可以直接发送文件至相关的软件中。可在工具栏的“发送”图标的下拉式菜单中，选择“Excel”。




注意: 您必须先先在电脑中安装 Excel 工作表软件, 才能直接发送辨识结果。并且在安装之后, 要先将之设定为发送模板, 才能在“发送”图标的下拉式菜单中直接选取套用。关于发送模板的设定, 请参考第 6 章的说明。

13. 辨识结果将会直接发送至 Excel 工作表软件中。

医院	地址	电话
1 上海市各大医院地址及联系电话一览表		
2		
3 医院	地址	电话
4 长宁区慢性防治院	仙霞路720号	64849181
5 卢湾区东南医院	淮海路1100号	63036583
6 黄浦区中医医院	北京西路130号	63277700
7 徐汇区龙华医院	双峰路450号	64339271
8 卢湾区妇幼保健院	延安中路393号	63272900
9 上海铁道医院	长阳路205号	65461434
10 上海第十人民医院	浦东老场镇大马路230号	59611047
11 上海市江湾医院	虹口区场中路240号	65422593
12 嘉定区妇幼保健院	嘉定区温南路39弄18号	59529069
13 徐汇区牙病防治所	肇嘉浜路685号	64037238
14 虹口区妇幼保健院	舟山路445号	65464235
15 闸北区市北医院	共和新路4500号	56816673
16 黄浦区疾病预防控制中心	福州路667号	63224295
17 上海浦南医院	浦东南路2400号	58391995
18 浦东新区公利医院	浦东新区苗圃路41号	58958730
19 南市区中内区综合医院	黄支路165号	63774871
20 朱家角人民医院	黄浦区朱家角镇	59240234
21 普陀区精神卫生中心	志丹路211号	56612948
22 杨浦区妇幼保健院	江浦路600号	65412226
23 上海宝钢十冶医院	宝山区丁家桥1号门外	66491174






在 Excel 工作表软件中的辨识结果

14. 您也可以选择保存文件。可在工具栏的“保存”图标  的下拉式菜单中，选择“保存本页辨识结果”。


保存文件辨识结果
保存本页辨识结果
保存本页原稿图片

15. 在保存对话框中，输入文件名称 GB_YP1，并指定存档类型为 EXCEL 文件格式，即可保存辨识结果。

保存本页辨识结果

保存在 (U):  Simplified    

文件名 (N):

保存类型 (T):  Excel 文件格式

将辨识结果保存成 EXCEL 文件格式。


第 10 章

自动辨识范例

丹青系统提供方便快速的自动辨识功能，从输入以至辨识等各项流程皆能自动执行，让您轻松地获得高正确率的辨识结果。

自动辨识文件

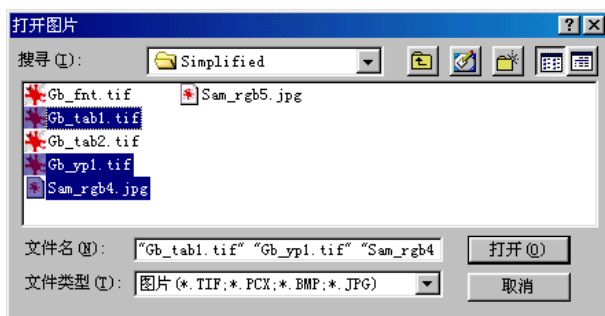
1. 选择“文件>新建”。

2. 按一下工具栏上的“自动”图标 。

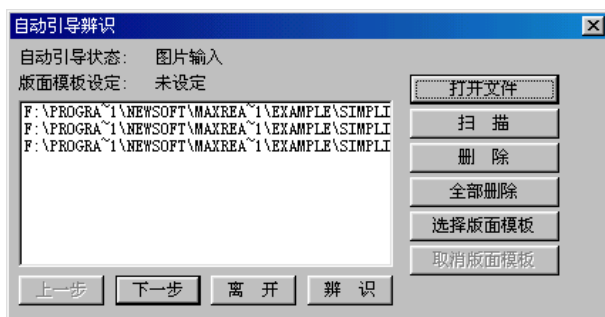
3. 在出现的对话框中，选择“打开文件”。



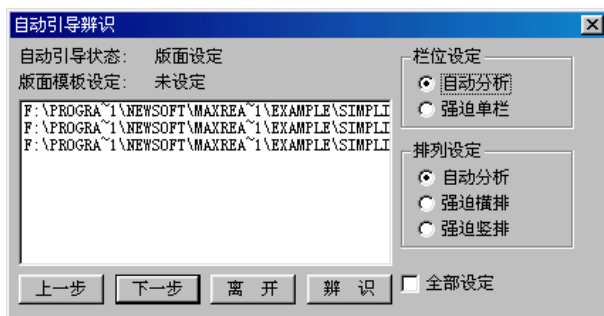
4. 打开丹青程序文件夹中 Example\Simplified 目录, 按住键盘上的“Ctrl”键并点选 Sam_RGB4. jpg、GB_TAB1.tif 及 GB_YP1.tif 等三个文件。



5. 按下“打开”，即可将欲辨识的文件读入。

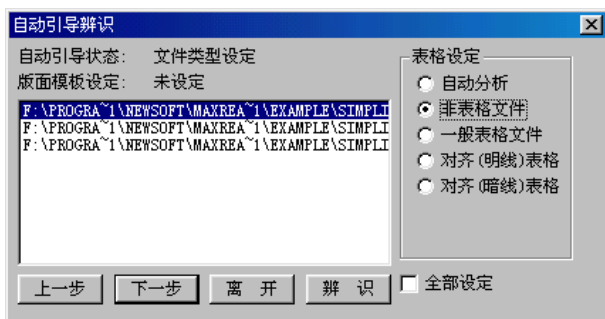


6. 选择“下一步”。
7. 在出现的选项中，分别点选各个文件，并一一设定为“栏位设定—自动分析”及“排列设定—自动分析”。



8. 选择“下一步”。

9. 在“表格设定”的选项中，点选 Sam_RGB4 文件名称并设定为“非表格文件”；选择 GB_TAB1 文件名称并设定为“一般表格文件”；以及选择 GB_YP1 文件名称并设定为“对齐(暗线)表格”。





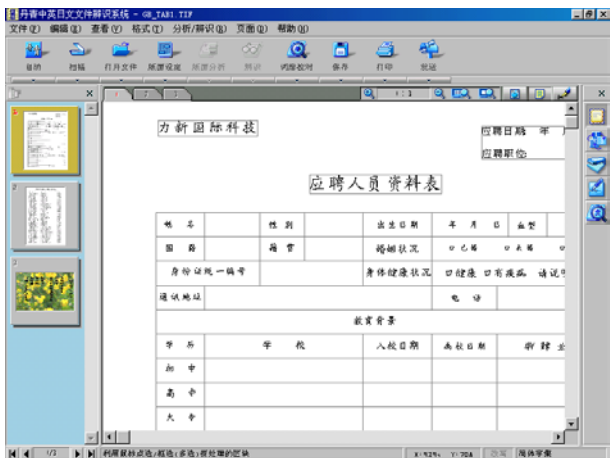
10. 选择“下一步”。

11. 在“文字设定”选项中，勾选“中文”、“英文”及“数字”。

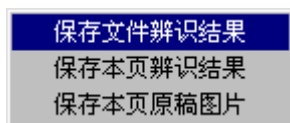


12. 按“下一步”，系统便开始执行三份文件的辨识工作。

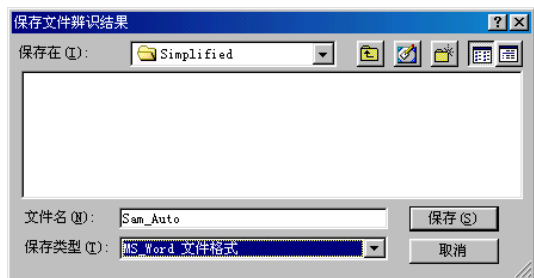
13. 选择查看工具箱上的“全页图文模式”及“文稿编辑模式”工具查看辨识结果，并依需要校对文稿。(校对文稿的详细步骤请参考第 5 章)。



14. 校对完成后，可在工具栏的“保存”图标下拉式菜单中选择“保存文件辨识结果”。



15. 在保存对话框中，输入文件名称为 Sam_Auto，并指定保存类型为 Word 文件格式。这三份文件将会被保存在同一个文件，并以换页符号区分各个文件。



保存文件辨识结果

附录 A

用语说明

用语	说明
区块属性	文件版面及资料内容的特性，如文字的横/竖排、文字的语言、是否含有表格、单/多栏位等。
剪贴板	文字及图片的暂存区块。用以保存“剪切”和“粘贴”命令处理的内容。
默认值	在程序中各种选项最初的设定。
解析度 (DPI)	解析度的度量单位。打印机和扫描仪的解析度是由每英寸所产生的点数来测量。DPI值愈高，解析度愈高。
内存	也称为 RAM（随机存取内存）。这是电脑暂时存放资料的区块。您可以将内存中的内容复制到硬盘或磁盘中永久保存。
下拉式菜单	当您选取菜单栏上的项目时出现的命令列表。
版面设定	文件资料内容的设定，如栏位设定、排列设定、表格设定、语言设定等。
版面分析	系统自动分析图片版面辨识区块、变更辨识区块顺序及设定文件属性
区块	文件图片上欲辨识的矩形区块
区块	版面分析后的矩形区块
分字	将相邻而辨识错的字元分开
合字	将相邻而辨识错的字元合并
分行	将因二行相连而辨识错的文字分开
合行	将被错误分割成二行的文字合并
区块分开再辨识	可分割被错误合并的区块
区块结合再辨识	可合并被错误分割的区块

候选字	与选取字字形上相似者或语意上前后相关者
学习新字	将常辨识错的字元输入到学习资料库中，增强辨识正确率

文件	
新建	打开新文件，退出原文件。
自动	自动输入、辨识及校对文件。
自动模板设定	设定自动功能的属性。
打开图片	打开图片。
保存文件辨识结果	保存整份文件的辨识结果，并以分页符号区隔各页面。
插入新页	在文件末端新增数页(由对话框选取)。
删除本页	将本页自文件中移除。
保存本页	以用户输入的名称保存本页图片或文字。若您选择以DOC、RTF、HTML格式保存文字，系统会保存文件内的所有内容(图形及文字)，并另将文件内所包含的图形依序编号保存。若您选择保存成纯文字档，则系统只会保存文件中的文字部分。
扫描文件	打开扫描界面（请参阅扫描仪的使用手册）。
扫描仪设定	选择扫描仪的来源。
打印	打印目前的文件。
设定打印机	选择打印机的来源。
发送	将文件图片或辨识结果直接发送至您事先设定的应用软件中。
发送模板设定	设定欲发送文件的应用软件及输出的文件格式。
系统设定	设定丹青系统程序的执行方式。
退出	离开丹青中英日文文件辨识系统。

编辑	
复原	复原前一项执行的动作。
剪切	将选取的区块剪切下来并存到系统的剪贴板上。
复制	复制选取的区块到剪贴板上。
粘贴	将剪贴板上的图片资料粘贴在打开的文件上。
清除	删除选取的区块。
全选	选择工作区内所有的图片或文字。
切除	切除不需要的图片部分。
反白	反转图片的文字颜色与背景颜色。
旋转	选择顺时针旋转90度、旋转180度、逆时针旋转90度或任意角度旋转等命令旋转工作区内的文件图片。
查找	查找文稿中的某个特定字词。
查找下一个	查找文稿中的下一个特定字词。
替换	以指定的字词替换某个特定字词。
查找第一个疑问字	查找目前页面中，丹青系统在辨识时所遇到的第一个疑问字。
查找下一个疑问字	查找下一个疑问字。

查看	
与窗口同宽	将屏幕上的图片放大或缩小至与窗口同宽。
全页显示	将屏幕上的图片以全页显示。
实际大小	将屏幕上的图片以实际大小显示。
缩小显示	以1/2到1/8的比例缩小显示屏幕上的图片。
放大显示	以2到8倍的比例放大显示屏幕上的图片。
放大	放大显示屏幕上的图片。
缩小	缩小显示屏幕上的图片。
原稿图片模式	显示输入的原稿图片。
全页图文模式	显示辨识后的图文版面。
文稿编辑模式	显示辨识后的文稿编辑模式。

工具栏	显示或隐藏工具栏。
缩略图窗口	显示或隐藏缩略图窗口。
编辑工具箱	显示或隐藏编辑工具箱。

格式	
版面设定	设定文件的资料属性及版面格式。
版面模板设定	设定常用文件的版面格式并可直接套用。
设定辨识字集	设定系统辨识时所使用的字集。
保存版面	保存目前工作区中的图片的版面格式资料，包括版面尺寸、设定的辨识区块及顺序等。
打开版面	套用已保存的版面格式。
放弃分析结果	放弃目前的版面分析结果，让用户重新设定，再行分析。
设成区块	设定框选部分为待辨识区块。
字体设定	指定文字区块的字体
字体大小设定	指定文字区块的字体大小
输入模式设定	设定文字输入的模式为“插入”或“改写”。

分析/辨识	
版面分析	分析图片上框选的辨识区块。
辨识	辨识分析后的文字。
词库校对	以指定的词库再辨识。
快速辨识模板设定	将常用的版面储存起来，所储存的模板会出现在识别下拉菜单中，可直接运用执行识别。
词库设定	设定个人常用的词库以作为词库校对的依据。
学习新字	让丹青系统在做文字辨识的时候，参考学习字库内的新字为辨识的根据。
删除学习字	删除学习字库中的学习字。
学习字库设定	指定学习字库作为辨识的依据。
区块再辨识	再次辨识所选择的区块。

放弃本页辨识	放弃目前的分割、辨识结果，让用户重新设定，再行分割、辨识。
放弃全部辨识	放弃所有文件的分割、辨识结果，让用户重新设定，再行分割、辨识。

页面	
第一页	显示文件中的第一页画面。
最后页	显示文件中的最后一页画面。
下一页	显示目前文件的下一页画面。
上一页	显示目前文件的上一页画面。
到第几页	显示指定的页面。
本页信息	显示本页图片的有关信息。

帮助	
帮助主题	选择帮助的内容或索引。
使用帮助	显示使用帮助的方法。
辅助提示	显示辅助的标题。
还原设定	还原至丹青系统程序的初始设定。
关于丹青	显示丹青程序信息、版本及版权。

工具栏及其下拉式菜单

工具栏中包含了在辨识过程里常用到的命令，其下拉式菜单中可列出与该图标相关的命令。您可以在各个图标的下拉式菜单中，按一下您要的命令，直接套用于文件上。

 <p>自动</p>	<p>自动</p> <p>自动输入、分析及辨识文件。您可选择“文件>自动模板设定”设定常用的自动模板；已设定的自动模板将会出现在下拉式菜单中，可供您直接执行自动功能。</p>
 <p>扫描</p>	<p>扫描</p> <p>从扫描仪输入图片。您可选择“文件>扫描仪设定”设定欲使用的扫描仪；所有已安装的扫描仪将会出现在下拉式菜单中。</p>
 <p>打开文件</p>	<p>打开文件</p> <p>打开欲辨识的图片。在其下拉式菜单中，会保留10组您最近打开过的文件。</p>
 <p>版面设定</p>	<p>版面设定</p> <p>设定文件的资料属性及版面格式。您可选择“格式>版面模板设定”设定常用的版面格式，如英文文件、通讯录、公文等；已设定的版面模板将会出现在下拉式菜单中，可供您直接使用。</p>

 <p>版面分析</p>	<p>版面分析</p> <p>自动分析图片上所框选的辨识区块。您可选择“格式>保存版面”，将常用的版面保存起来。在下拉式菜单中，会保留10组您最近保存的版面文件。</p>
 <p>辨识</p>	<p>辨识</p> <p>执行辨识命令。您可选择“分析/辨识>快速辨识模板设定”，将常用的版面保存起来；您所保存的版面会出现在下拉式菜单中，可供您直接套用执行辨识。</p>
 <p>词库校对</p>	<p>词库校对</p> <p>以指定的词库再辨识。您可选择“分析/辨识>词库设定”，将常用的词库保存起来。系统保留10组词库供您设定。这10组词库会出现在下拉式菜单中，您可直接选择数个词库，再执行辨识。</p>
 <p>保存</p>	<p>保存</p> <p>以用户输入的名称保存本页图片、文字或整份文件。您可选择“保存文件辨识结果”来保存整份文件的辨识结果，“保存本页辨识结果”来保存本页的辨识结果，或是“保存本页原稿图片”来保存原稿的图片。</p>
 <p>打印</p>	<p>打印</p> <p>打印目前的文件。您可选择“文件>打印机设定”设定欲使用的打印机；所有已安装的打印机将会出现在下拉式菜单中。</p>
 <p>发送</p>	<p>发送</p> <p>将文件图片或辨识结果直接发送至您所安装的电子邮件软件中。您可选择“文件>发送模板设定”设定欲发送的应用软件，如文字处理器或图片处理软件等；您所设定的发送模板会出现在下拉式菜单中，可让您直接点选将文件传至该软件。</p>

编辑工具箱图标

	放大显示 放大显示屏幕上的图片。
	缩小显示 缩小显示屏幕上的图片。
	全屏显示 将屏幕上的图片以全屏显示。
	与窗口同宽显示 将屏幕上的图片放大或缩小至与窗口同宽。
	原稿图片模式 显示输入的原稿图片。
	全屏图文模式 显示辨识后的图文版面。
	文稿编辑模式 显示辨识后的文稿编辑模式。

	<p>设定辨识区块</p> <p>框选待辨识区块。使用鼠标在图片上拖动框选要做辨识的区块。</p>
	<p>变更辨识区块顺序</p> <p>每一个框选出来的区块都有一个辨识的序号，您可以决定它们的先后顺序。</p> <p>(请参考变更辨识区块顺序的使用帮助。)</p>
	<p>选取图片区块</p> <p>可使用鼠标在图片上拖动框选图片区块，之后再执行“切除”或“设定区块”等功能。</p>
	<p>橡皮擦</p> <p>清除文件上的杂点，以提高辨识效率。</p>
	<p>绘笔工具</p> <p>添补图片漏白的部分。</p>
	<p>区块结合再辨识工具</p> <p>可合并被错误分割的区块，再次辨识。</p>
	<p>区块分开再辨识工具</p> <p>可分割被错误合并的区块，再次辨识。</p>
	<p>文字校对工具</p> <p>显示辨识时的疑问字。</p>
	<p>文字合并再辨识工具</p>

	将相邻二个或数个辨识错的字元合并并重新辨识。
	文字切割再辨识工具 将相邻二个或数个辨识错的字元分开并重新辨识。
	文字行合并再辨识工具 将被错误分割成二行的文字合并并予以重新辨识。
	文字行切割再辨识工具 将因二行相连而辨识错的文字分开并予以重新辨识。

扫描的建议

当您扫描一份文件时，原稿品质的好与坏会直接影响到扫描的结果。而当您在扫描不同的文件图片时，所使用的解析度值也跟着不同。通常扫描一般文件（文字高度约为 3mm）时，建议您使用 400dpi 的解析度值。若字体稍小，扫描时则建议您提高解析度。

为了方便系统辨识，请您在使用系统辨识文件以前，用橡皮擦工具擦去图片上的杂点，以提高系统的辨识效率。

如何改善辨识品质

影响辨识品质的因素有三：图片品质、扫描时的解析度(dpi)、扫描时的明亮度。

- 图片品质：扫描时文件要放端正(误差以三度为上限)、图片要清晰。
- 扫描时的解析度(dpi)：一般而言，图片内文字的大小在 40~50 Pixel 时，丹青会有最佳的表现，过与不及，皆会降低辨识品质。
- 以中国时报及联合报本文字体的大小(3mm)为例，可用 400dpi 来扫描；较大的字体建议您使用 300dpi，详细资料请参考图例一“扫描解析度建议”。
- 扫描时的明亮度：太淡（造成断线--请参考图例四）或太浓（糊成一团--请参考图例三）的图片都会降低系统的辨识率。因此在调整扫描的明亮度时，应注意不要让笔划简单的字（如中、大、口）的笔划断掉，也不要让笔划复杂的字（如团、丽、吁）的内部糊成一团。

若两者无法兼顾时应以保留笔划简单的字的横向笔划为优先。

图例一：扫描解析度建议

建议值：400dpi

力新国际科技股份有限公司 (10 point)

力新国际科技股份有限公司 (11 point)

力新国际科技股份有限公司 (12 point)

建议值：300dpi

力新国际科技股份有限公司 (12 point)

力新国际科技股份有限公司 (14 point)

力新国际科技股份有限公司 (16 point)

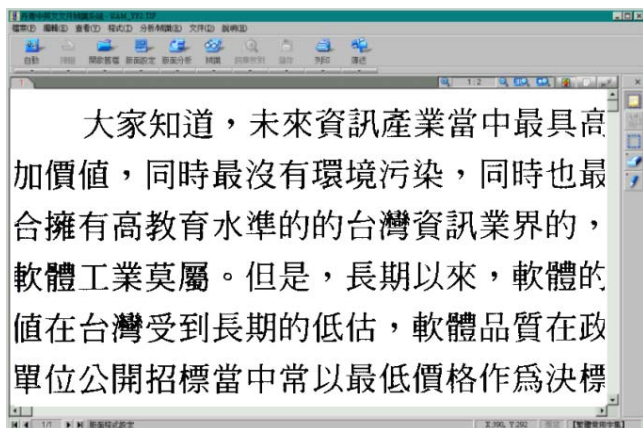
建议值：200dpi

力新国际科技股份有限公司 (18 point)

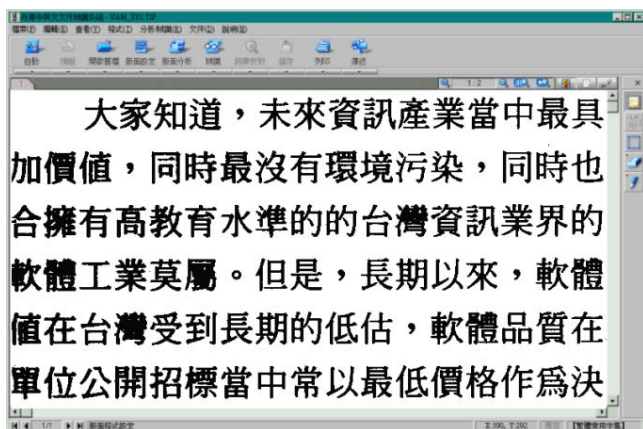
力新国际科技股份有限公司 (20 point)

力新国际科技股份有限公司 (22 point)

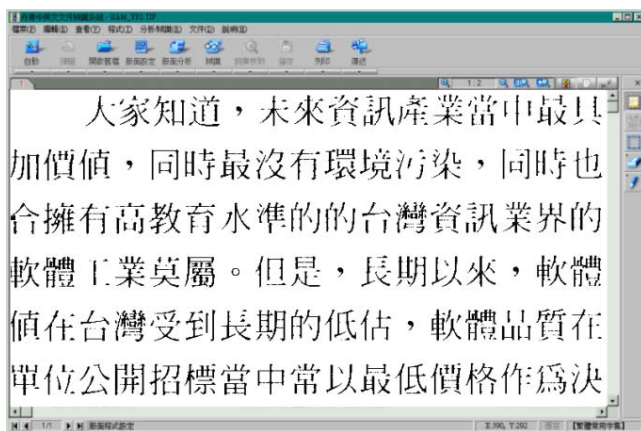
图例二：标准



图例三：太浓



图例四：太淡



就实际例子而言，如果文件本身为白底黑字，则可利用扫描仪驱动程序上的“自动明亮度调整”按钮来为你设定扫描图片的明亮度值；若是文件为黄底黑字(如报纸类)，则可再增加亮度值，若是效果依旧不好，可再增加数个单位的值。

总体来说，当您觉得丹青系统的辨识品质未达到您的需求时，可试着调整图片的解析度与明亮度。其中明亮度是个经验值，可能各家扫描仪而不同。另外，试着给予辨识区块属性，如设定区块属性(有/无)汉字、设定区块属性(有/无)英文字母以及设定区块属性(有/无)数字等，也有助于系统的辨识。