OCR:光学字符识别技术

所谓 OCR （Optical Character Recognition 光学字符识别）技术，是指电子设备（例如扫描仪或数码相机）检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后
用字符识别方法将形状翻译成计算机文字的过程；即，对文本资料进行扫描，然后对图像文件进行分析处理，获取文字及版面信息的过程。由于 OCR 是一门与识别率拔河的技
术，因此如何除错或利用辅助信息提高识别正确率，是 OCR 最重要的课题，ICR（Intelligent Character Recognition）的名词也因此而产生。而根据文字资料存在的媒体介质不
同，及取得这些资料的方式不同，就衍生出各式各样、各种不同的应用。

编辑本段 OCR 的发展史

要谈 OCR 的发展，早在 60、70 年代，世界各国就开始有 OCR 的研究，而研究的初期，多以文字的识别方法研究为主，且识别的文字仅为 0 至 9 的数字。以同样拥有方块文字的日本
为例，1960 年左右开始研究 OCR 的基本识别理论，初期以数字为对象，直至 1965 至 1970 年之间开始有一些简单的产品，如印刷文字的邮政编码识别系统，识别邮件上的邮政编码，
帮助邮局作区域分信的作业；也因此至今邮政编码一直是各国所倡导的地址书写方式。 OCR 可以说是一种不确定的技术研究，正确率就像是一个无穷趋近函数，知道其趋近值

，却只能靠近而无法达到，永远在与 100%作拉锯战。因为其牵扯的因素太多了，书写者的习惯或文件印刷品质、扫描仪的扫描品质、识别的方法、学习及测试的样本……等等，多
少都会影响其正确率，也因此， OCR 的产品除了需有一个强有力的识别核心外，产品的操作使用方便性、所提供的除错功能及方法，亦是决定产品好坏的重要因素。 一个 OCR
识别系统，其目的很简单，只是要把影像作一个转换，使影像内的图形继续保存、有表格则表格内资料及影像内的文字，一律变成计算机文字，使能达到影像资料的储存量减少、
识别出的文字可再使用及分析，当然也可节省因键盘输入的人力与时间。 从影像到结果输出，须经过影像输入、影像前处理、文字特征抽取、比对识别、最后经人工校正将
认错的文字更正，将结果输出。

影像输入：

欲经过 OCR 处理的标的物须透过光学仪器，如影像扫描仪、传真机或任何摄影器材，将影像转入计算机。科技的进步，扫描仪等的输入装置已制作的愈来愈精致，轻薄短小、
品质也高，对 OCR 有相当大的帮助，扫描仪的分辨率使影像更清晰、扫除速度更增进 OCR 处理的效率。 影像前处理：影像前处理是 OCR 系统中，须解决问题最多的一个模块，从
得到一个不是黑就是白的二值化影像，或灰阶、彩色的影像，到独立出一个个的文字影像的过程，都属于影像前处理。包含了影像正规化、

去除噪声、影像矫正等的影像处理，及
图文分析、文字行与字分离的文件前处理。在影像处理方面，在学理及技术方面都已达成熟阶段，因此在市面上或网站上有不少可用的链接库；在文件前处理方面，则凭各家本领
了；影像须先将图片、表格及文字区域分离出来，甚至可将文章的编排方向、文章的提纲及内容主体区分开，而文字的大小及文字的字体亦可如原始文件一样的判断出来。
文字特征抽取：单以识别率而言，特征抽取可说是 OCR 的核心，用什么特征、怎么抽取，直接影响识别的好坏，也所以在 OCR 研究初期，特征抽取的研究报告特别的多。而特征可
说是识别的筹码，简易的区分可分为两类：一为统计的特征，如文字区域内的黑/白点数比，当文字区分成好几个区域时，这一个个区域黑/白点数比之联合，就成了空间的一个数
值向量，在比对时，基本的数学理论就足以应付了。而另一类特征为结构的特征，如文字影像细线化后，取得字的笔划端点、交叉点之数量及位置，或以笔划段为特征，配合特殊
的比对方法，进行比对，市面上的线上手写输入软件的识别方法多以此种结构的方法为主。　　对比数据库：当输入文字算完特征后，不管是用统计或结构的特征，都须有一比
对数据库或特征数据库来进行比对，数据库的内容应包含所有欲识别的字集文字，根据与输入文字一样的特征抽取方法所得的特征群组。
对比识别：

这是可充分发挥数学运算理论的一个模块，根据不同的特征特性，选用不同的数学距离函数，较有名的比对方法有，欧式空间的比对方法、松弛比对法（Relaxation）、动态
程序比对法（Dynamic Programming，DP），以及类神经网络的数据库建立及比对、HMM（Hidden Markov Model）…等著名的方法，为了使识别的结果更稳定，也有所谓的专家系
统（Experts System）被提出，利用各种特征比对方法的相异互补性，使识别出的结果，其信心度特别的高。　　字词后处理：由于 OCR 的识别率并无法达到百分之百，或想加强
比对的正确性及信心值，一些除错或甚至帮忙更正的功能，也成为 OCR 系统中必要的一个模块。字词后处理就是一例，利用比对后的识别文字与其可能的相似候选字群中，根据前
后的识别文字找出最合乎逻辑的词，做更正的功能。　　字词数据库：为字词后处理所建立的词库。

人工校正：

OCR 最后的关卡，在此之前，使用者可能只是拿支鼠标，跟着软件设计的节奏操作或仅是观看，而在此有可能须特别花使用者的精神及时间，去更正甚至找寻可能是 OCR 出错的
地方。一个好的 OCR 软件，除了有一个稳定的影像处理及识别核心，以降低错误率外，人工校正的操作流程及其功能，亦影响 OCR 的处理效率，因此，文字影像与识别文字的对照，
及其屏幕信息摆放的位置、还有每一识别文字的候选字功能、拒认字

的功能、及字词后处理后特意标示出可能有问题的字词，都是为使用者设计尽量少使用键盘的一种功能，当然
，不是说系统没显示出的文字就一定正确，就像完全由键盘输入的工作人员也会有出错的时候，这时要重新校正一次或能允许些许的错，就完全看使用单位的需求了。

结果输出：

其实输出是件简单的事,但却须看使用者用 OCR 到底为了什么？有人只要文本文件作部份文字的再使用之用，所以只要一般的文字文件、有人要漂漂亮亮的和输入文件一模一
样，所以有原文重现的功能、有人注重表格内的文字，所以要和 Excel 等软件结合。无论怎么变化，都只是输出档案格式的变化而已。如果需要还原成原文一样格式，则在识别后
，需要人工排版，耗时耗力。

编辑本段中文 OCR

光学符号识别技术是一种汉字文稿的自动输入方式，它通过光学扫描仪和计算机的配合，经 OCR 软件将图像数据进行运算分类后，将图像数据转化为计算机内码，可以极大减
轻数据录入工作的强度，提高数据录入的速度。 文献资料的数字化录入，一般分为： 1、纯图像方式。 2、目录文本、正文图像方式。 3、全文本方式。
4、全文索引方式。文本方式和图像方式的混合体。 我国在 OCR 技术方面的研究工作起步较晚，在 70 年代才开始对数字、英文字母

及符号的识别进行研究，70 年代末开始进行

汉字识别的研究，到 1986 年汉字识别的研究进入一个实质性的阶段，不少研究单位相继推出了中文 OCR 产品. 我国目前使用的文本型 OCR 软件主要有清华文通 TH-OCR、赛酷 OCR

、北信 BI-OCR、中自 ICR、沈阳自动化所 SY-OCR、北京曙光公司 NI-OCR（已被中自汉王并购）等，匹配的扫描仪则使用市面上的平板扫描仪。

编辑本段ＯＣＲ衡量标准

衡量一个 OCR 系统性能好坏的主要指标有：拒识率、误识率、识别速度、用户界面的友好性，产品的稳定性，易用性及可行性等方面。

编辑本段ＯＣＲ工作原理

识别过程: 书本级：中文，英文；简体，繁体； 版式级：竖排，横排；有无分栏； 行切分 字切分 识别：真正的 OCR 识别过程，图像信息还原成文本

信息 后处理：人工干预，主要集中在前四个阶段。 识别精度可以达到 99%

编辑本段 OCR 识别率决定因素

1.图片的质量，一般建议 150dpi 以上 2.颜色，一般对彩色识别很差，黑白的图片较高，因此建议 ocr 的为黑白 tif 格式 3.最重要的就是字体，如果是手写识别率很低

。 国内 OCR 识别简体差错率为万分之三，如果要求更高的精度需要投入更大的人工干预。繁体识别由于繁体字库的不统一性（民国

时期的字库和现在繁体字库不统一），导致
识别困难，在人工干预下，精度能达到 90%以上（图文清晰情况下）。OCR 是计算机输入技术的一种，它通过模式识别将文字的图像文件转化为可编辑的文本文件，彻底改变
了计算机纸介质资料输入的概念。只要用扫描仪将文本图像输入计算机，就可转化为可修改的文本文件，这比手工输入速度快了几十倍。随着 OCR 技术的广泛应用，它正逐渐被人
们所知晓。国际软件巨头微软在研发 XP 系统的时候，就意识到 OCR 的市场需求，在发布的 Office 2003 中全面配装了 TH-OCR(北京文通信息技术有限公司开发）;硬件方面的领袖企业
英特尔公司也确定 TH-OCR 为 MMX 技术支持项目。　　近期，一些大公司意识到 OCR 的好处，开始在自己的产品中捆绑 OCR 技术。Google 已经启动 OCR 软件的开发工作，在它的招聘启
示中这样写道：“Google currently "reads" almost every web page in the world. Come help us read all the printed material as well!”(Google 现在已经能够“阅读
”世界上几乎所有网页，你的到来将让 Google 阅读所有印刷信息!)。随着 google 启动 OCR 开发工作，OCR 应用进入了全面爆发时代。无论是让计算机对文字进行排版输出，还
是要让计算机认识它看到文字，所有这一切都是为我们生活服务。信息化和数字化的进程，让我们不再安于用十指敲击键盘来输入数据。人们希望能将时间和精力投入到更具创造

性的工作中去，因而希望计算机等辅助设备能更具智慧。OCR(Optical Character Recognition，光学字符识别)技术就是其中的一项，跟打印技术相对，它是让计算机认字的一种
技术，这远比打印复杂得多。　　经济竞争带来更多的商务活动，每个活动上名片都是必不可少的主角，名片的管理产品也应运而生，名片识别管理工具同样也是以 OCR 技术为核
心的产品。通过名片识别工具将名片进行扫描、识别、分类，不仅能够导入手机、PDA 等，而且还能为名片信息进行备份，不用担心遗失。文通 e-card 就是一款优秀的名片识别管
理产品，OCR 技术能把商务生活打理得有条不紊，节约更多的时间。现在，几乎所有的扫描仪和一体机上都配装 OCR 软件，比如 HP、UNISCAN、EPSON、CANON、 LENOVO 等扫描仪厂商
捆绑的就是文通 TH-OCR。