第1章 绪 论

1.1 引言

人类社会已进入了信息时代，尤为重要的标志之一是互联网的发展已经深入人们的生活，从宽度、广度和深度方方面面改变了和改变着人们的生活方式，也改变了世界。信息化使得信息的获取、传输、交换和使用成为影响社会发展的重要因素，信息事业的发展极大地影响了国家的发达和民族的兴旺，也因此得到世界各国的极大关注。

在计算机信息化迅速发展的过程中，信息的电子化处理已成为一种不可逆转的趋势，需要解决如何把大量的已产生或将产生的印刷或手写的海量文档信息高效地输入计算机这样的问题，即使在未来，这也是必不可少的一步。

将电子化文档输出为纸质文档，激光照排技术带来了对历史上铅与火排版技术的革命，使信息化得到重要发展。但反之，要将无处不在、无时不有的介质上的印刷或手书文档，自动变成计算机可以阅读（查询和检索等）的电子文档，却是十分重要，但却相当难以实现的。虽然可以采用人工键入的方法，但完全无法满足信息化时代对高速、大数据和大容量的需求。

如何满足全球信息化对于文档数字化高速、大数据、大容量的急迫需求，利用计算机模式识别技术进行文字和文档的自动识别，实现形形色色的文档的自动电子化，为计算机信息化发展打下坚实的基础是我们研究工作的目的，也是本书写作的动因。

《文字识别：原理、方法和实践》一书源于自 20 世纪 80 年代开始作者对汉字识别的研究和探索，以及 30 余年持续的研发和产业化工作，因此有必要对这些研究工作加以总结和汇总。

《文字识别：原理、方法和实践》的写作基本上沿着模式识别与文字和文档的信息化这两条线索展开。

第 1 条线索是模式识别，是本书的理论依据。由于文字识别是最典型的，也是目前最有成效的模式识别技术，因此我们有必要首先介绍模式识别以及解决模式识别问题的统计模式识别的基本理论和方法，从提出模式识别信息熵理论开始，包括模式识别特征提取、特征选择和压缩、分类器设计、上下文相关识别方法等基本问题的研究探讨。

第 2 条线索是文字和文档的信息化，这是本书的中心内容。文字是信息的最集中表现，汉字记载了 5 000 余年中国的历史和现代文明的发展。尤其是在计算机信息化时代，文字信息化是信息化时代的基础问题也是关键的问题，特别是困难的文档信息的计算机自动输入问题。在西方文字信息化已取得较完善发展的 20 世纪 60—70 年代，数量巨大、结构复杂的汉字信息化却遇到汉字计算机输入的特殊困难，成为汉字计算机信息化的拦路虎。完善解决多种文字和文档自动识别计算机输入等问题，是本书研讨的主要内容，包括利用统计模式识别方法，对多文种文档识别的众多关键问题进行较为详细的研究和探讨，等等。

本书介绍了文字和文档识别的理论、方法和实践应用。根据模仿人类视觉模型，提出有别于结构分析的基于文字图像的统计模式识别方法，有效突破了汉字输入计算机对信息化的壁垒，取得了文字识别令人瞩目的进展。从模式识别信息熵的分析说明了统计模式识别方法的理论基础，分析了从文字图像中提取识别特征的方法，以及文字识别中分类器的学习和设计方法；提出汉字的综合识别研究，以及文本识别必须解决的版面分析、文字切分和利用上下文识别后处理等重要问题，最后，总结了文字识别研究的重要进展情况并对未来工作加以展望。

1.2 文字和汉字

文字是人类社会文明的基石，是人类信息最重要的载体，文字信息是信息最集中的表现，是人类信息传承、交换、记载的依据。应当说，人类文明源于文字的出现，人类文明的发展更离不开文字。在信息化时代的今天，尤其是在互联网全球化之时，文字信息数字化对于人类文明发展更具特殊的意义。这种无所不在和无处不有的海量大数据文字信息的数字化要求，注定了文字识别的不可或缺及其在世界范围内广泛的应用需求。

文字是语言的符号表示，世界上使用的文字基本上可以分为以下几种：

拉丁字母、基里尔字母、阿拉伯字母、印度字母、汉字系统及其他(韩语、蒙古语、希伯来语等)文字等。

汉字是世界上最古老的三大文字系统之一。其他如古埃及的圣书字、两河流域苏美尔人的楔形文字已经失传,仅有唯一的中国的汉字沿用至今。

汉字,是中国人创造的意音文字书写系统,也是当今世界上唯一仍被广泛采用的意音文字和独源文字,推估历史可追溯至约 4 000 年前的夏商时期。汉字主要用于书面记录汉语(因而又可称为中文),一个字对应汉语的一个音节和一个语素;也用于记录日语、朝鲜语(韩语)和古代越南语等东亚、东南亚多种语言,文字性质与中文不尽相同。

秦始皇统一中国后,统一了中国的文字。“书同文”的历史从此开始。文字的统一有力地促进了不同民族间的文化传播,对中国的统一以及东亚各国的文化交流发挥了重要作用,为世界文字史所罕见。

汉字的特点有:字根组字(以有意义的 869 个声母及 265 个形母的象形字为字根组成各种汉字)、表意、书同文、兼容并蓄等。以基本的象形指事字为基础,发展了形声、会意的组字法,以组合方式,细化大量的字出来,使得文书上的记载越来越精密,到今天一直成为造字的主力。汉字由一个或以上的字根以二维方式(欧语系是一维文字)在特定的空间、配置在一个正方块内而组成,因此有方块字的别称。

汉字是以意念的表达需要,组合所需字根部件于一个方块中,合成千千万万的字。每一个汉字或字根,由横、竖、撇、捺、拐、点等基本笔画构建而成,笔画数目从最少一个笔画到 36 个笔画之多,可见汉字笔画结构的复杂程度变化之大。

而汉字的构造分为单字、部件、笔画和笔段 4 个层次。单个汉字是一个由笔画构成,结构完整、具有意义和读音的二维图形,是形、音、义的统一体。我们读书认字,就是根据字形而知其音、识其义。用计算机自动识别汉字也是这个意思。

从语义表达的层次,有字、词、短语、句和篇章之分。

1.2.1　文字的代码表示

为解决文字和汉字信息的相互正确交换、存储、传输以及共享,作为文字信息的计算机处理的基础,国际上和我们国家都陆续出台和制定了一系列文字和汉字的字符集与标准代码,即对某一个符号或汉字的内涵所赋予的代码表示。文字的机内编码标准是重要的国际和国家的信息化标准。美

国在 20 世纪 60 年代就已发展和制定了英文的字元集和交换码，以及美国的国家标准 ASCII 编码(Standard Code for Information Interchange)，对每一个字符或符号用一个字节编码，并进一步演变为世界性的电脑字元编码标准 ISO 646 和 Unicode。

由于全球信息化发展的要求，1990 年国际标准化组织 ISO(International Organization for Standardization)颁布了国际语言文字统一编码标准 ISO 10646(简称 UCS-4)，是 4 字节的字符编码标准，包括世界主要语言文字的统一编码，其已发表的标准包括有 70 205 个汉字。

我国汉字的国标码(《中华人民共和国国家字符标准》，简称 GB 码)机内编码国家标准有：

1980 年发布的 GB 2312，它规定了 6 763 个简体汉字的编码，其中包括 3 755 个一级汉字，3 008 个二级汉字。一级汉字的使用频度达到 99.99%。

1993 年发布的 GB 13000，又称为 GBK 标准，它规定了包括 20 902 个简繁体汉字和韩文、日文在内的 CJK 字符编码，以及藏、维、蒙等民族文字。

2000 年发布的 GB 18030，它规定了包括 GB 13000 字符在内的以及扩展的 6 582 个古汉字，总计有 27 484 个汉字编码；最近还将扩展 4 万余字，总数达到接近 7 万余字。

汉字编码还包括：

- Big 5 码，收录 13 053 个汉字，包括在台湾和香港使用的繁体汉字。
- Unicode，简称 UCS-2，是国外一些计算机厂商提出并推广的一种可容纳世界各国语言文字的统一编码体系，每字符 2 字节。汉字字符集包括 2 万余汉字。

我们可以看到，具有成千上万巨大字符集是汉字有别于其他文字的突出特点。

1.2.2 汉字的字体字形

远古时代的汉字是一种象形文字，是模仿事物形状而刻画的图案。殷商时代的甲骨文和金文虽仍保留若干象形图案，但已包含一些表意图形；结构上也由独体字发展而成合体字，并出现很多形声字。春秋战国时通用的文字是大篆和小篆，秦代因“奏事繁多，篆书难成，隶人(指胥吏)佐书，曰隶书。”篆书笔画圆转，隶书笔画方折，便于书写。使用隶书，基本上改变了原有汉字的体形，奠定了楷书的基础。汉初为使汉字书写更为方便，出现了

“草隶”，及至草书和行书。草书笔画潦草，往往难以辨认，取而代之的是楷书。由于楷书形状方正，笔画平直，又名正书或正楷，魏晋以后楷书成为汉字的正宗，一直到现在，仍然是汉字的楷模。

汉字是象形文字，早期的汉字图形并不都是方形的，楷书成为正宗之后，汉字才成为名副其实的方块字。尤其是在印刷体汉字出现之后，每个汉字的大小相同，长宽相等，成为汉字的重要特征之一。

如图 1.1 所示，汉字的基本字体包括：篆、隶、楷、行、草。图中名称以绿色标示的，是历史发展的字体，表示了汉字字形的历史发展过程；以红色标示的，则是书法或美术设计上的字形。图中还包括书法和印刷使用的美术字体，前者如欧体、颜体，后者如宋体、黑体。而简体汉字出现在楷书、行书之后，本无所谓的简化隶书、草书，图中所列[隶体]，仅为模仿隶书风格借书法美术而建模写出来的简化汉字而已。

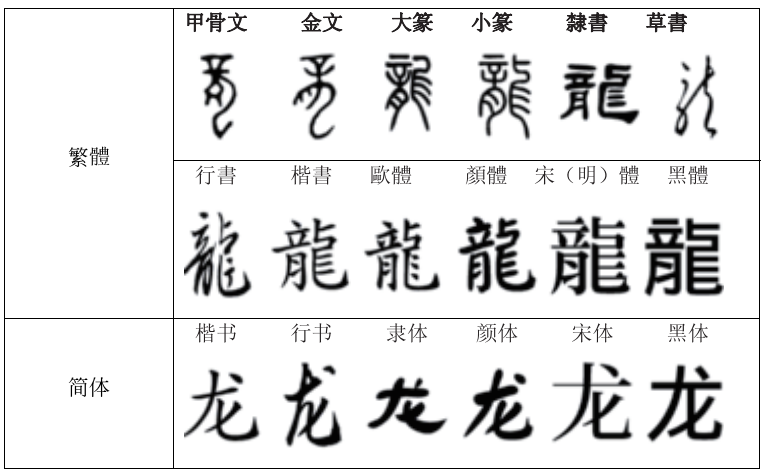

图 1.1　汉字字体一览表(见彩插)

印刷术发明之后，产生了便于印刷的宋体字，结构方正，笔画横细竖粗，便于刻字，又易于活字排版。元、明两代出现的元体和明体字基本上与宋体字相同，统称为宋体。20 世纪初出现仿宋体，其结构与宋体相同，只是横竖粗细基本相同，以后又有笔画粗而黑的黑体字出现。近几十年来，宋体、黑体、仿宋体和楷体，已成为我国印刷品汉字的主要字体，近年来为排版美观等需求，还发展了其多种变体。由于计算机的推广使用，计算机生成了多种字体且其变形层出不穷，其目的是使字形更美观，但其字形基本上是围绕着

基本字体而变化的。然而变形字体的层出不穷，也为汉字的识别带来一定的困难。

汉字中宋、仿、黑、楷字体等变形字体图形多达 199 种，部分示例如图 1.2 所示。

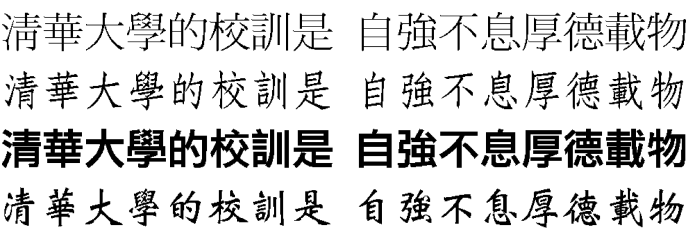

图 1.2　宋、仿、黑、楷字体图形

汉字字形分为繁体和简体，具有不同的编码，相当于不同的汉字。

汉字的大小尺寸变化也是汉字重要的形状特征之一。印刷体汉字的大小通常用不同的字号表示：字形从小到大发生变化，从最小号字到特大号字顺序为七号汉字、小六号、六号、小五号、五号、小四号、四号、三号、小二号、二号、一号、小初、初、小特、特直到特大号汉字。从最小的七号汉字到最大的特大号汉字，字形大小变化了近 10 倍，以适应不同排版和阅读的需要。一般在文字识别中，对于字号的变化，经过对字符图像大小尺寸的规一化，即可基本消除字号变化对字符识别的影响。

1.2.3　汉字的特点

汉字的首要特点是数量巨大，编码为 GBS 2312 的简体一级汉字有 3 755 个，二级汉字有 3 008 个；一级和二级简体汉字总计为 6 763 个；繁体汉字以 Big 5 码收录的有 13 053 个。如果扩大汉字编码和应用的范围，GB 18030 全部汉字数量已经达到 4 万～7 万字，是世界上具有最大字符集数量的文字。显然，巨大数量的汉字字符集给汉字识别带来的困难也是巨大的，使汉字识别成为超多类模式识别的困难问题。近来往往会出现简繁体汉字共用或简繁体汉字混用的情况，从汉字识别的角度看，这相当于增加了汉字识别的字符类别数，更增加了汉字识别的困难。

汉字的另一个重要特点是，汉字是由复杂的笔画结构构成的，因此，复杂的笔画结构是汉字的基础特征，也是汉字的本质特点，不同汉字的复杂程度极不相同，最简单的汉字仅一个笔画构成，如“一”，最复杂的汉字可达 36

个笔画之多；从结构模式分析的角度来看，复杂模式结构确实为汉字的结构识别算法带来不小的负担，但是从汉字的结构统计算法来说，复杂的汉字结构往往增加了汉字之间的差异性，汉字识别反而从中获得益处，使汉字获得优于其他文字的识别性能。

汉字的复杂笔画结构可以分层分解为由笔段、笔画、字根、单字 4 个层次组成，如果考虑到词是词义表达的最小单位，则可以增加语义层次为 5 个层次。构成了汉字基本笔画的汉字不同层次结构，这可以为汉字的结构分析和汉字结构识别带来很大的益处。

汉字使用的频度也是汉字的重要特点。虽然汉字的数量极其巨大，但其利用频度极不相同，且使用频度极高的汉字数量十分有限，GB 2312 一级 3 755 个汉字的使用频度高达 99.99%，日常生活常用汉字仅 2 000 余字。

综上分析，我们可以看到，汉字不仅数量极其浩大，汉字字符达数千至数万(4 000～70 000)之多；字符结构非常繁杂，汉字的笔画数最多可达到 36 画；字形变化巨大：由于字体的不同，给印刷体汉字识别带来一定的困难；而更困难的是无约束的手写汉字的识别，由于书写者不同、书写条件不同，使得汉字字形变化差异多样。可想而知，类别数量巨大、字符结构复杂、字形无约束巨大变化，给超多类高性能汉字识别带来了巨大的挑战。

汉字识别的困难主要表现在结构复杂和变化、数量巨大的字符识别上，而汉字的复杂结构却又为汉字识别提供了足够的汉字特征信息，使识别的困难得以化解。而且汉字较规则和聚团的方块字形，也为汉字文本的切分带来很大便利。和英文等其他文字相比，其字符数目虽然很少，但笔画简单、结构信息的缺乏不仅给识别带来困难，而且字形的不规则也给字符切分带来巨大的困难，成为文本识别难以克服的障碍。实际上，目前汉字文档(无论是印刷或手写的)识别性能已获得优于其他文字文档识别的性能。

汉字不仅有识别的优势，而且汉字是最精练和高效的文字。著名学者季羡林说“汉字是世界语言里最精炼的一个语种。同样表达一个意思，如果英语需要 60 秒，汉语 5 秒就够了。”而表示同样内容的英文文本的英语字母数与汉语文本中汉字字符数目之比平均可达到 3.25 之多。同时，汉字具有极强的组词功能，通过少量的常用汉字，可以生成大量新的词条和词语。而英语需要学习的新词汇达到 1 000 万条，因此汉字是最具扩展学习能力的文字。这些优点为汉语文化的发扬光大打下坚实的基础。

1.2.4 中文信息处理

中文信息处理指的是对汉字及其他民族文字的计算机信息化处理，即用计算机对汉语（包括口语和书面语）进行转换、传输、存储、分析等信息处理的科学，是我国信息化发展的基础。显然，中文信息计算机处理必须要解决好汉字的输入、存储、传递、输出等问题。北京大学王选教授激光照排的创新，解决了汉字的计算机输出问题，极大地推动了中文信息处理的发展。但是，中文信息处理还必须解决中文的计算机输入问题。由于西方研发的打字机键盘适用于西文的键盘输入，利用键盘输入巨大数量的汉字困难重重。在中文信息处理发展的初期，由于汉字输入遇到的极大困难，曾引起汉字能否适应计算机时代的极大困惑和争论，甚至曾引发了“汉字拉丁化”的思潮。

随着上千种中文键盘输入法的出现，主要包括表音输入和表形输入方法，或两者兼之，使得利用键盘的汉字的计算机输入方法得以推广使用，成为解决汉字计算机输入的基本手段。但由于手工键入的繁琐和低效，完全无法满足和适应海量大数据资源和高速信息化要求。汉字和文档的自动识别、汉字的语音识别输入等技术的发展和日趋成熟，为汉字计算机输入带来新的希望。汉字识别和汉字语音识别成为 20 世纪早期研究者努力攀登的高峰。通过 30 余年的研究探索，汉字及重要民族文字文档的自动识别已经成功实现，并得到广泛推广和应用。而且，在目前汉字识别计算机输入水平的情况下，已经超过了一般拼音文字识别的计算机输入水平，这对汉字信息化的发展产生了巨大的推动力，使千年古老的汉字能在当今计算机信息化时代重放光芒。

1.3 文字识别和汉字识别

文字是人类信息最集中的表象和最重要的载体，对人类文明的传承和发展起着决定性的作用。在计算机信息化过程中，在互联网深入改变了世界和人们的生活方式的今天，各种文字记录都迫切面临着“电子化”的要求，以期利于计算机处理、通信、检索和转换。西方国家在 20 世纪中期开始研究和发展西文光学字符识别（optical character recognition，OCR）技术和文档识别技术，以使大量文字资料能快速、方便、省时省力和及时地自动输入计算机，实现信息处理的“电子化”。显然，汉字的信息化处理也将大大依赖

于汉字识别和语音识别的发展。

我们知道，人们认字的过程是根据对文字的字符图像的视觉观测，借助大脑的认知，对文字的类别加以区分辨识的过程，而不受字符图像千变万化的影响（无论是印刷的、手写的，还是摄像获取的）。

什么是文字识别？就是要使计算机实现人们通过大脑完成的识图认字的功能。也就是利用计算机将人们可以阅读的文字图像信息，自动转化为计算机可以阅读、可查询的以计算机内码表示的文本信息。

文字识别系统就是基于对文字图像的传感输入，利用计算机完成对文字图像内涵的文字类别的模式辨识，并将文字的类别以字符编码表示和输出的系统。也就是说，文字识别就是对观测的文字图像内涵的文字类别的模式辨识和转换，而与文字的大小、字体、印刷字的字模、不同个人书写的变化和差异等均无关系。

我国是使用汉字的国家，汉字记载了我国五千年的悠久历史文明，而且在现代文明中起着不可替代的作用。但是，数量浩大、结构繁杂、变化多端的汉字难以输入计算机的问题，曾一度成为汉字信息化发展的拦路虎。在汉字信息化的过程中，众多汉字编码与汉字键盘输入方案（主要有字形编码和拼音编码两类）都是拆分汉字以适应为西方文字设计的键盘输入，费时费力。寻找自动和快速的汉字文档计算机输入方法成为人们深思和努力求解的问题。因此，研究和发展汉字识别的理论和方法，解决数量浩大、结构繁杂、变化多端的汉字识别问题，并解决好汉字文本资料、手写汉字文本、手写汉字以及手写数字等海量文档的自动、快速、方便地输入计算机这类问题，对于汉字的信息化具有特殊重要的意义。对于汉字识别的研究，也是关于计算机智能感知和认知问题的研究，所关注的是如何利用计算机实现人类的智能感知和认知的研究。毫无疑问，这是当前模式识别和计算机视觉学科的重要课题，具有极其重要的理论和实际意义。

国际上，1966 年 IBM 公司的 Casey 和 Nagy 首次发表了汉字识别的文章[1]，国内的汉字识别研究开始于 20 世纪 70 年代末，我国科学工作者经过近 30 年的研究和努力，已经从理论和实践上基本解决了汉字识别问题，即：实现了对各种实际文本图像的计算机自动识图认字，用计算机自动实现对各种文字，包括古今中外（简繁汉、英、日、韩、藏、维哈柯、阿等）多种文种的、各种印刷字体的、各种复杂图文版面的识别、理解和重构，不仅解决了印刷文本的识别问题，而且还解决了手写（包括联机手写和脱机手写）汉字和数

字的识别问题。识别系统在国民经济各行各业得到普遍推广和使用，成为国家信息化不可或缺的手段。

总结起来，文字识别就是利用计算机将纸张上（或其他物理器件上）人们可以阅读的文字图像信息，自动转化为计算机可以阅读和查询的以计算机内码表示的文本信息。

这种文字信息的数字化过程是现代信息化时代的基础，是使得计算机能够对各种文本信息进行信息的智能利用、检索和查询的前提条件。而文字识别作为文字信息高质量和高效率自动数字化的基本手段，在现代信息化时代的重要性就可想而知了。

1.4 文字识别研究历程

文字识别技术的研究已经有半个多世纪的历史。参考 Arica 在文献[2]中的划分方法，我们可将字符识别的研究历程大致分为 3 个阶段。

（1）早期阶段（20 世纪 50—70 年代）：字符识别的研究出现在计算机诞生之后不久，最初起步于对印刷体字符的识别，50 年代中期出现了相应产品[3]，随后逐渐地扩展到手写字符识别，可识别的字符集也从简单的数字、英文扩展到其他各种文字。1966 年，IBM 公司的 Nagy 等人首次发表了关于汉字识别的文献[1]。这个时期的字符识别方法受到计算机运算能力和数据采集水平的极大限制，以简单的图像匹配为主，识别性能低，对字符图像的质量也有着很严格的要求。

（2）理论发展阶段（20 世纪 80—90 年代中期）：这是字符识别的实验室研究空前活跃的一个时期，计算机运算速度的提高和模式识别理论的成熟共同促进了字符识别技术的迅速发展，世界各地的学者们掀起了字符识别研究的热潮，每年均有大量研究文献问世，各种各样的方法被应用于字符识别中来[4-9]。与此同时，真正实用化的识别系统也开始进入市场。在汉字识别方面，日本学者首先在特征匹配方法上取得了重要的进展[6,7]，大大提高了汉字的识别率，而国内学者也不甘落后，以中国科学院自动化所、清华大学电子系、北京邮电大学为代表的研究单位先后致力于汉字识别的研究，并很快在识别性能上取得了长足进步，达到了国际领先水平。

（3）全面应用阶段（20 世纪 90 年代末期到现在）：进入 90 年代末期后，大规模的字符识别研究热潮有所减退，新的识别方法出现得不多，但是随着