

汉字识别的 系统与集成

戴汝为 郝红卫 肖旭红

智能化丛书

智能自动化丛书

汉字识别的 系统与集成

戴汝为 郝红卫 肖旭红

浙江科学技术出版社

THE SERIES OF INTELLIGENT AUTOMATION

**SYSTEM AND INTERGRATION OF
CHINESE CHARACTER
RECOGNITION**

Dai Ruwei, Hao Hongwei, Xiao Xuhong

**ZHEJIANG SCIENCE & TECHNOLOGY
PUBLISHING HOUSE**

智能自动化丛书

汉字识别的系统与集成

戴汝为 郝红卫 肖旭红

*

浙江科学技术出版社出版

浙江印刷集团公司印刷

浙江省新华书店发行

开本:850×1168 1/32

印张:11 插页:4

字数:220 000

1998年12月第一版

1998年12月第一次印刷

ISBN 7-5341-0883-7/TP·44

定 价:20.00 元

责任编辑:褚天福 莫沈茗

封面设计:孙 菁

Published by Zhejiang Science and

Technology Publishing House

347 Tiyyuchang Road, Hangzhou, China

©1998 by Dai Ru Wei etc.

First published in 1998

Printed in Zhejiang Printing Group

ISBN 7-5341-0883-7/TP·44

内 容 简 介

本书对“计算机汉字识别”这一模式识别的重要分支进行了介绍,并简要讨论了模式识别与形象思维的联系,目的在于从人的模式识别中得到一些启发。在扼要介绍了模式识别的一些数学方法之后,以脱机手写汉字识别、联机手写汉字识别以及计算机笔迹鉴别作为主要线索,以讨论“从定性到定量的综合集成法”的构思为主导思想,通过人机结合的途径,从汉字识别的方法、系统以及特征提取等3个不同的层次进行集成,从而取长补短、互相补充,得到集成型汉字识别系统。本书内容新颖,一方面反映了近年来汉字识别方面的部分研究成果;另一方面也包括了一些与汉字识别密切相关的内容,如样本库、版面理解、预处理与后处理等。本书对于从事汉字识别、模式识别、人工智能以及思维科学等领域的研究人员、科技人员具有一定的参考价值。

ABSTRACT

The present book introduces computer recognition of Chinese characters which is an important branch of pattern recognition. The relationship between pattern recognition and thinking in image is also briefly discussed in order to draw inspiration from pattern recognition of human being. After the general mathematical methods of pattern recognition are briefly introduced, the problems in off-line handwritten Chinese character recognition, on-line handwritten Chinese character recognition and writer identification system are discussed. The emphases are placed on integration of method level, system level and feature level, so that their complementary can be applied to the design of integrated Chinese character recognition system. Directed by the idea of metasynthesis from qualitative to quantitative approach, the integrated Chinese character recognition systems are obtained by means of human-machine integration. The contents of the book are new and original. The book covers not only some recent research results, but also several related topics such as database, layout understanding, preprocessing and postprocessing. It is valuable for researchers and graduate students in areas of Chinese character recognition, pattern recognition, artificial intelligence and noetic sciences, etc. .

《智能自动化丛书》编委会

名誉主编 宋 健

主 编 戴汝为

编委(按姓氏笔画为序)

于景元	王 珏	石青云	冯纯伯
边肇祺	朱剑英	李国杰	李衍达
吴启迪	何新贵	宋国宁	张 侃
张 钹	陆汝铃	赵沁平	姜 桐
袁保宗	席裕庚	韩京清	路甬祥
熊范纶	潘云鹤	戴冠中	瞿寿德

Editorial Committee

Honor Editor-in-Chief

Song Jian

Editor-in-Chief

Dai Ruwei

Members of Editorial Committee

Yu Jingyuan	Wang Jue
Shi Qingyun	Feng Chunbo
Bian Zhaoqi	Zhu Jianying
Li Guojie	Li Yanda
Wu Qidi	He Xingui
Song Guoning	Zhang Kan
Zhang Bo	Lu Ruqian
Zhao Qinqing	Jiang Tong
Yuan Baozong	Xi Yugeng
Han Jingqing	Lu Yongxiang
Xiong Fanlun	Pan Yunhe
Dai Guanzhong	Qu Shoude

出版说明

当前我们正面临着一场信息革命,在这场革命中计算机扮演着重要的角色,而现代化通讯、人工智能与多媒体技术等科学技术的发展将对这场革命起着十分重要的推动作用。目前,科技界已有一种共识:采用人工智能原理和方法研制智能系统,以使传统的自动化走向智能自动化,是自动化学科发展的必然趋势。智能自动化受到国内外的重视,我国在这方面也已开展了许多基础及应用研究。

为了推广和交流智能自动化的研究成果,在宋健教授的关怀下,以戴汝为教授为主编,我们组织出版了这套“智能自动化丛书”。

近期内,本丛书计划针对国内外有关智能自动化的最新研究进展及我国专家学者取得的成果,在智能系统的综合集成、智能控制、人工神经网络原理及应用、智能信息处理及汉字识别等方面出版一系列内容新颖的专著。

我们希望这套丛书的出版,能为我国智能自动化的发展作出一点贡献。

浙江科学技术出版社

1998年4月

前 言

自从 20 世纪 40 年代数字计算机问世以来,以微电子、计算机、通信为主要支柱的信息技术得到了突飞猛进的发展,从而将人类从工业化社会向信息化社会推进。最初的计算机是用来进行科学计算的,但随着信息技术的发展,计算机渗入到了各个领域,在各行各业中都能大显神通,起着极其重要的作用。由于在使用计算机的时候,总需要把数据或资料送入计算机,或者把需要解决的问题以计算机能够接受的方式送入计算机,因而如何把人与计算机两者连接起来,也就是如何建立人与计算机两者之间的“接口”,就成为必须加以解决的问题。

在人类的历史发展过程中,人与人之间最初是用简单的图形进行通信与交流的,以后采用象形文字,直至现在用以符号为基本单元的符号串所表示的语言和文字。计算机问世的初期,数据和程序是通过键盘被送入计算机的,后来人们很自然地考虑到能否采用人类所习惯的通信和交流方式与计算机打交道。为此,首先考虑的是计算机能不能对语音、文字和图形有认识能力,只有在能够认识的前提下,才可能以此为基础进行进一步的处理。

对于使用汉字与汉语的中国人来说,用汉语与汉字同计算机进行交互是一项具有现实意义的工作。因为,我们所处理的文件资料,大部分是用印刷的或手写的汉字完成的。要把我国建设成现代化的强国,不可能光靠人力来处理信息,而必须依靠信息技术来进行信息处理。本书将主要从两方面来探讨汉字识别问题,以及更为广泛的模式识别问题。

一方面是从如何能有效地与计算机进行交互的角度,来讨论计算机识别(recognizing)模式的问题。我国的邻邦日本,由于比较早地认识到了信息和知识是一种能够创造财富的资源,一种比自然资源更为重要而宝贵的资源,所以对信息技术给予了高度重视。70年代初,为了适应计算机技术竞争的需要,从长远观点解决人机交互问题,使计算机的输入输出与机器的运算速度能迅速适应,日本通产省实施了一个发展信息技术的“模式信息处理计划(Pattern Information Processing System Project)”,简称 PIPS 计划。该计划的核心在于发展模式识别技术。其目的是对文字、声音、图像、景物等加以处理、分析、分类、描述及理解,解决计算机与人类活动的环境直接通信的问题,使计算机能自动识别字符、图形及声音信号等,从而能把这种类型的信息直接送入机器,或机器直接输出这种类型的信息。1978 年在京都举行的第 4

届国际模式识别大会上,日本的专家介绍了他们进行 PIPS 计划所取得的成就,并组织会议的参加者到东京的邮局,参观邮政信函自动分拣系统。这种高度自动化的系统采用 10 个阿拉伯数字进行邮区编码,通过当时十分先进的手写数字识别技术对信函自动分拣。这是字符识别技术的有效应用,使大家对这一新技术产生了深刻的印象。当时,日本学者致力于研制识别 800 个日本常用汉字的工作。近 15 年来,PC 机的出现和推广应用,为研究工作提供了方便而有效的工具。为数众多的年轻人,尤其是一些攻读硕士学位或博士学位的年轻学者,参与了模式识别及字符识别的研究工作,对类别十分庞大的汉字分类问题进行了较深入的探讨与研究。目前,我国内地、香港、台湾及东南亚地区的学者研究用计算机对联机及脱机手写汉字进行识别时,大都着眼的是解决工程技术问题。结果表明,计算机能识别汉字,说明计算机表现出具有一定的“智能行为”,有点像人一样“聪明”。但实际的情况是:计算机识别手写汉字与人脑识别手写汉字的机理完全是两回事。尽管不同人书写的汉字变化多端、龙飞凤舞,但人脑能够正确地加以识别。而目前计算机进行汉字识别所用的方法往往是以设计者的“技巧”为主。它虽然能达到较高的识别率且速度很快,但毕竟相当“死板”。

另一方面是从智能系统的角度来探讨。当人们看到计算机借助编制的程序,能够认识图形,能够听懂声音,并按照语音命令作出相应的反应,会感到惊奇。在家喻户晓的《天方夜谭》中有个小故事:面对着一个关闭的石洞,只要喊一声“芝麻开门”,两扇石门就会自动打开;如果喊的不是“芝麻”而是其他植物的名字,石洞的门就不会打开。而当前,用模式识别技术来实现把门打开的这种声控系统就并不是天方夜谭,是不难办到的事。这样的系统看起来颇有点像人那样,具有在不同情况下能做出相应反应的能力,所以称声控系统是一种“智能系统(intelligent system)”。

汉字识别及模式识别的研究也属于智能系统的研究范畴。对智能系统的研究可以从不同的途径进行。90年代我国著名科学家钱学森提炼出开放的复杂巨系统的概念(开放的复杂巨系统问题,如社会环境、地理环境、社会问题、人体问题、人脑问题),并且提出处理开放的复杂巨系统的方法是从定性到定量的综合集成法(metasyntesis)。这一方法以实践论为主立足点,强调人的主导作用及经验的重要性。这个方法的最初表达为:处理开放的复杂巨系统的方法是“将专家群体(各方面有关的专家)、数据和各种信息与计算机技术有机地结合起来,把各种学科的科学理论结合起来,这三者本身也构成一个系统。这个方

法的成功应用就在于发挥这个系统的整体优势和综合优势”。我国有“集大成”的说法,就是说,把一类非常复杂的事物的各方面综合起来,集其大成。综合集成法实质上是综合集成从定性认识到整体定量认识的方法,简称“大成智慧工程”。从而我国开始形成一门以“开放的复杂巨系统”为主要内容的“复杂巨系统”科学。综合集成法,其核心是强调人与计算机两者相结合,这对研制智能系统(包括模式识别系统),具有指导性的作用。

在这种构思的影响下,本书的作者及中国科学院自动化研究所人工智能实验室的一些科技人员,从模式识别的方法、识别系统的互补以及特征抽取等不同的层次,对如何进行集成进行了比较系统的探讨,其目的在于研制综合性能好的汉字识别系统。通过这些工作,改变了传统的有关模式识别是实现从特征空间到类别空间的一个非线性映射的观点(由于引入反馈机制,可以把模式识别看成为一种非线性动态过程)。这种观点曾经由系统科学专家 H. Haken 首先提出,但他的论述未得到进一步的实现和发展。我们则以人工神经网络为工具,构成若干分类系统为例,用有教师的学习对网络进行训练,然后用相类似的人工神经网络对分类系统的结果进行集成。在集成网络的结构方面,由于引入反馈,因而形成了闭路系统,并进一步

用有教师的学习对集成网络加以训练。可以看出,在整个系统中,各分类系统及集成系统中参数的确定,都有人参与,从而把人的经验与知识注入到系统之中,从整个系统的结构与运行方面,论述了模式识别系统是一个非线性动态系统。这样一来,就把模式识别系统与强调反馈作用的控制系统从理论上、结构上统一了起来,形成了“集成型模式识别”的研究方向。

应该指出,在这个方向上还有许多工作有待完成。本书将对这方面的学术观点与研究结果,作一扼要的介绍。如果书中的内容能引起大家的兴趣,那么就一定程度上实现了著者的愿望。

著 者

1998. 2

目 录

第一章 概论	[1]
1.1 汉字识别与模式识别	[1]
1.2 模式识别与形象思维	[8]
1.3 监督学习	[14]
1.4 集成型汉字识别	[19]
参考文献	[27]
 第二章 模式识别的一些数学方法	 [29]
2.1 线性判别函数	[30]
2.2 最小距离法	[37]
2.3 最近邻法	[39]
2.4 人工神经网络	[41]
2.5 句法方法	[53]
2.6 语义句法方法	[69]
参考文献	[86]
 第三章 样本库	 [88]
3.1 引言	[88]

3.2 国外样本库简介	[92]
3.3 脱机手写汉字样本库	[97]
3.4 联机手写汉字样本库	[102]
3.5 印刷体汉字样本库	[103]
3.6 自由手写数字样本库	[104]
3.7 汉字识别系统的性能指标	[105]
参考文献	[110]
 第四章 版面理解	 [112]
4.1 引言	[112]
4.2 版面的结构分类和语言描述	[116]
4.3 版面理解的算法实现	[129]
4.4 行字分割	[150]
参考文献	[153]
 第五章 预处理与后处理	 [155]
5.1 脱机手写汉字识别系统的构成	[155]
5.2 预处理算法	[158]
5.3 后处理	[172]
参考文献	[176]
 第六章 脱机手写汉字识别	 [178]
6.1 引言	[178]