
人机自然交互

Human-Computer Nature Interaction

张有为 等著



国防工业出版社

人机自然交互

Human-Computer Nature Interaction

张有为 等著

国防工业出版社

·北京·

图书在版编目(CIP)数据

人机自然交互/张有为等著. —北京:国防工业出版社, 2004. 9

ISBN 7-118-03544-0

I. 人… II. 张… III. 人一机系统—研究
IV. TB18

中国版本图书馆 CIP 数据核字(2004)第 065391 号

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号)

(邮政编码 100044)

北京奥隆印刷厂印刷

新华书店经售

*

开本 710×960 1/16 印张 17 $\frac{1}{4}$ 314 千字

2004 年 9 月第 1 版 2004 年 9 月北京第 1 次印刷

印数:1—3000 册 定价:58.00 元

(本书如有印装错误,我社负责调换)

致 读 者

本书由国防科技图书出版基金资助出版。

国防科技图书出版工作是国防科技事业的一个重要方面。优秀的国防科技图书既是国防科技成果的一部分,又是国防科技水平的重要标志。为了促进国防科技和武器装备建设事业的发展,加强社会主义物质文明和精神文明建设,培养优秀科技人才,确保国防科技优秀图书的出版,原国防科工委于1988年初决定每年拨出专款,设立国防科技图书出版基金,成立评审委员会,扶持、审定出版国防科技优秀图书。

国防科技图书出版基金资助的对象是:

1. 在国防科学技术领域中,学术水平高,内容有创见,在学科上居领先地位的基础科学理论图书;在工程技术理论方面有突破的应用科学专著。
2. 学术思想新颖,内容具体、实用,对国防科技和武器装备发展具有较大推动作用的专著;密切结合国防现代化和武器装备现代化需要的高新技术内容的专著。
3. 有重要发展前景和有重大开拓使用价值,密切结合国防现代化和武器装备现代化需要的新工艺、新材料内容的专著。
4. 填补目前我国科技领域空白并具有军事应用前景的薄弱学科和边缘学科的科技图书。

国防科技图书出版基金评审委员会在总装备部的领导下开展工作,负责掌握出版基金的使用方向,评审受理的图书选题,决定资助的图书选题和资助金额,以及决定中断或取消资助等。经评审给予资助的图书,由总装备部国防工业出版社列选出版。

国防科技事业已经取得了举世瞩目的成就。国防科技图书承担着记载和弘扬这些成就,积累和传播科技知识的使命。在改革开放的新形势下,原国防科工委率先设立出版基金,扶持出版科技图书,这是一项具有深远意义的创举。此举势必促

使国防科技图书的出版随着国防科技事业的发展更加兴旺。

设立出版基金是一件新生事物,是对出版工作的一项改革。因而,评审工作需要不断地摸索、认真地总结和及时地改进,这样,才能使有限的基金发挥出巨大的效能。评审工作更需要国防科技和武器装备建设战线广大科技工作者、专家、教授,以及社会各界朋友的热情支持。

让我们携起手来,为祖国昌盛、科技腾飞、出版繁荣而共同奋斗!

**国防科技图书出版基金
评审委员会**

国防科技图书出版基金 第四届评审委员会组成人员

名誉主任委员	陈达植				
顾问	黄宁				
主任委员	刘成海				
副主任委员	王峰	张涵信	张又栋		
秘书长	张又栋				
副秘书长	彭华良	蔡镛			
委员	于景元	王小谟	甘茂治	冯允成	
(按姓名笔画排序)	刘世参	杨星豪	李德毅	吴有生	
	何新贵	佟玉民	宋家树	张立同	
	张鸿元	陈火旺	侯正明	常显奇	
	崔尔杰	韩祖南	舒长胜		

前 言

自然科学中有 2 个重要的概念,就是能量与信息。能量这个概念在科学技术中所起的支配作用是人们所熟知的,而与能量的存在具有同样悠久历史的信息的概念,在科学上取得应有的地位,却只有半个世纪。400 年前、200 年前、50 年前分别出现了制造业、动力业、信息业。20 世纪 40 年代末期,信息科学的开拓者美国学者香农、维纳和苏联学者柯罗莫格洛夫所发表的经典著作标志着信息科学发展的起点。

信息科学的一个重要研究方向是人机交互,近几年来已趋于活跃。人与人的交流是很自然的,这是一个人人交互过程,是人的思想通过语言和情感表达与理解的高级生命双向交互过程。人是作为最高级生命而存在的,其他生命间的交互即使存在,也是处于低级形态,主要表现在同类同种同属的动物之间。生命之间的交互是通过听觉、视觉、触觉、嗅觉和躯体姿态来完成的。对于最高级生命——人类来说,语言是思想表达的主要方式,其信息是通过听觉与视觉通道来传递的。所谓人机自然交互是指人与机器的自然交互或人与计算机的自然交互。自然交互区别于人工的交互,是通过赋予机器的听觉和视觉自然地完成的。智能化的实现主要通过计算机,或者说智能机器的智能是由它包含的计算机而赋予的。

人类的进步,其重要特征是制造工具,智能化的工具就是智能化的机器。人们渴望机器能受人支配,听从人的指使,只要人动嘴要它做什么,它都能够理解,并按要求去做。同时机器还能把对于人的表述的理解反馈回来,回答人提出的问题,达到交互的目的。人类对机器的要求还不止于此,甚至要使人与机器的对话达到人与人之间交流那种自然的水平。这是除人与人之间以外,其他非高级生命与人类交流无法或难以达到的。这种想法所实现的是生命与非生命之间的对话,涉及非生命对生命思想的理解。虽然非生命机器的思维都是人这种高级生命注入给它,并经过训练和它的自学习而完成的,或许相当于人类接受教育的过程。

自然信源是极其丰富的,遂以人类的进步、生态繁茂和经济繁荣的方式向前发展。然而,人造信息体制与自然信息体制是不同的,不但自然信源尚待发掘,二者之间的翻译与接口界面也尚待开发。人机自然交互的研究是发展人造信息与自然信息的接口,本质上是改造人造信息渐与自然信息体制相兼容。人机自然交互作为信息科学与技术的一个先进研究领域,必然涉及哲学、人类学、生物学、数学、物

理学、电子科学、计算机科学、人工智能、信息论、控制论、认知科学、心理学以及伦理学,是它们互相渗透、互相结合的产物。

人机自然交互正如人与人交互一样,首先应能识别交互对象,即知道在与谁交互对话;其次要理解对话的内容;第3要理解对方的情态,即是兴奋、喜悦、无动于衷、烦恼、愤怒等情感;第4要能对说话人进行定位与跟踪,好像人与人交流过程中总能注视对方,有多个人说话时总能注视正在说话者或区分说话人。在人机交互中存在一个双向过程,人对于机器的理解相对来说是易于完成的,而难点在于机器对人的理解。所以对于人机自然交互的研究,即研究生命与非生命对话过程中,将主要研究非生命对于生命形象、语言和情感的理解。

人类的语言本质上是双模态的,音频语言与说话人发出的声波有关,视频语言与说话人嘴唇、舌头和面部肌肉的运动有关。音频与视频所对应的正是人的听觉与视觉。人机自然交互的研究就是要赋予机器听觉与视觉的功能,能识别人并能理解人的说话内容和感情,进而进行推理与思维,执行人对它的要求或与人进行交流。

本书共有6章,第1章是导论;第2章、第3章将分别从视觉和听觉角度研究对说话人的识别;第3章、第4章是从听觉和视觉角度研究对说话内容的识别与理解;第5章研究听觉和视觉双模态的融合问题;第6章介绍对于人机交互研究的最基础的条件:听觉-视觉双模态语言识别数据库。关于情态识别问题、对说话人定位与跟踪问题以及多模态人机交互网络环境问题,本书除在导论中给予介绍外,不再另立章节论述。这是由于情态识别问题在理论和技术上都还不很成熟,还没有达到著书的可能;而对说话人定位与跟踪问题就技术实现上来说有很多书籍都可参用。关于多模态人机交互网络环境问题,虽然我们进行了大量研究,但是由于本书篇幅所限,不再列为一章研究和讨论。本书用大量的篇幅着重介绍了作者在实验室中进行的研究结果和验证。然而人机自然交互是一个广阔的研究领域,由于本书主要是就作者所得的研究结果进行论述,只能说是研究了人机自然交互的若干问题。为了弥补这种不足,本书采用极简洁的文字介绍了国内外的主要理论和方法,同时给出了详尽的参考文献,便于需要深入了解的读者查阅。

信息科学是一个年轻的自然科学分支,而人机自然交互是更为年轻的一个研究方向,它被人们所重视还是近10年的事,从严格的科学意义上来说,它离完善还有相当的路程,当前仍然处于初始阶段。我们虽然在本书中从听觉-视觉的角度来研究问题、建立模型,但是我们还没有做到建立一个从识别说话人、跟踪说话人、识别理解说话内容和情感的统一模型,尽管我们相信它是存在的。科学的终极在于提供一个简洁的理论去描述整个人机自然交互过程,看来这是非常困难的。世界一切事物存在一个“不变”的规律,就是“变”,科学研究过程亦是如此。对于科学工作者来说,变就是进步。人类已经可以观测到100亿光年以外的宇宙,也可以看到

亿分之一纳米以下的微观结构,不断向认知的极限挺进,人们也一定可以认识人机自然交互的规律与本质。

人机自然交互的实现将会给机器带来革命性的变化,当它应用于电子产品、通信设施、机械设备、交通工具、人工智能、智能仪器、多媒体、情报采集、身份认证、安全防范以及武器现代化时,将会对科学技术、生产领域、国家安全、社会的工作方式和生活方式等方面产生深远的影响。

本书由张有为主撰,甘俊英、何强、蒙山、应自炉和张歆奕对本书中涉及的一些理论和进行了实验验证和开拓,并进行了相应章节的撰写工作(他们撰写的章节是:第2章和第4章甘俊英,第3章何强,3.9.3节张歆奕,第5章蒙山,第6章应自炉。他们的排名是依姓氏的汉语拼音第1个字母为序排列的,他们对本书做出了同样重要的努力,应视为并列的第2作者)。应自炉进行了本书的计算机编排工作。

本书在写作过程中参考和引用了国内外许多作者的有关论述和结果,从中得到了启发和教益,也得到了学术界朋友的支持与鼓励。中国科学院院士清华大学李衍达教授、中国科学院院士中国科学院声学研究所侯朝焕教授,在2001年底本书初稿形成后,进行了评审,给予了鼓励并推荐本书出版。中国电子学会信号处理分会主任委员北京交通大学博士生导师袁保宗教授、中国电子学会信号处理分会副主任委员北京航空航天大学博士生导师毛士艺教授、中国科学院声学研究所博士生导师杜利民教授对于我们在本书涉及人机自然交互技术和信号与信息处理研究方法等方面曾多次给予指导。在我们进行人机自然交互领域的研究中,得到了2项“863”计划子课题的支持,和3项广东省自然科学基金(No. 960631, No. 000872, No. 032356)的支持,同时也得到了广东省教育厅和五邑大学的大力支持。在此一并表示衷心的感谢。感谢本书全体作者的母校——北京航空航天大学培育之恩。感谢国防科技图书出版基金评审委员会对于本书出版的有力支持。

在本书的一些章节我们使用了 ORL, CAVSR 1.0, CAVBSR-WUIIS(1.0)3 个数据库的人脸资料,用以表示对人脸图像的实验处理结果,为本书增加了所述内容的说服力。在进行人机自然交互的研究过程中,我们参阅了众多的论文和著作,有些列入了参考文献,有些由于篇幅所限尚未列入,我们从中获得了许多启发和教益,在此谨表谢忱。研究生赵向阳、何元烈、王东、戴伟、张莉、汪晓东、黄生、傅筠、周延蕾等都进行了许多有意义的研究和开发工作,他们的研究成果丰富了人机自然交互的某些研究。

由于作者水平所限,本书难免有错误与不当之处,欢迎批评指正。

目 录

第 1 章 导论	1
1.1 从人机交互到人机自然交互	1
1.1.1 人机交互和人机自然交互	1
1.1.2 人机自然交互的主要功能与特征	2
1.1.3 人机自然交互发展的社会与科学技术背景	3
1.2 自然信源与人造信息的接口界面	3
1.2.1 自然信源	3
1.2.2 人造信息	4
1.2.3 进一步的思考	5
1.3 识别交互对象	6
1.3.1 说话人识别	6
1.3.2 说话人识别的途径	7
1.4 识别交互内容	11
1.4.1 识别交互内容是交互中的核心问题	11
1.4.2 对自然语言的理解	15
1.5 听觉-视觉双模态融合	16
1.5.1 融合问题	16
1.5.2 融合策略	16
1.5.3 融合策略与识别算法	17
1.6 对人类情态的感知	18
1.6.1 听觉-视觉双模态情态识别问题	18
1.6.2 显性信道和隐性信道	19
1.7 多模态网络环境、定位跟踪和数据库	20
1.7.1 人机自然交互系统	20
1.7.2 网络环境	21
1.7.3 对说话人的定位与跟踪	22
1.7.4 双模态数据库	23
1.8 人机自然交互带来的生产方式、工作方式和生活方式的变革	25

1.8.1	人机自然交互的实现将引发变革	25
1.8.2	军事上的应用及民用前景	25
第2章	视觉——说话人识别与人脸识别	27
2.1	说话人识别问题	27
2.2	人脸图像的预处理	29
2.2.1	人脸图像的检测与定位	30
2.2.2	人脸图像的标准化	30
2.3	人脸图像的特征提取与识别	37
2.3.1	几何特征法	38
2.3.2	特征脸法和局部特征法	39
2.3.3	弹性模型法	39
2.3.4	神经网络法	40
2.3.5	不变矩特征法	40
2.4	人脸特征自适应主元提取法	42
2.4.1	统计主元分析法	43
2.4.2	自适应主元提取法	43
2.4.3	自适应主元提取法的收敛性分析	45
2.4.4	应用实例	49
2.5	人脸图像奇异值特征提取法	55
2.5.1	奇异值特征	55
2.5.2	奇异值降维压缩	56
2.5.3	应用实例	57
2.6	最佳鉴别向量特征提取法	63
2.6.1	核函数 Fisher 鉴别	63
2.6.2	广义核函数 Fisher 最佳鉴别	66
2.7	人脸识别图像分层算法及应用实例	72
2.7.1	用于人脸识别的人脸图像分层算法	73
2.7.2	用于人脸识别的人脸图像快速分层算法	80
第3章	听觉——说话人识别、语音识别与理解	86
3.1	语音识别问题	86
3.1.1	语音识别技术的发展	86
3.1.2	语音识别系统的分类	89
3.1.3	语音识别系统的基本构成	89
3.2	语音信号的特征	90
3.2.1	语音信号的数字化	90

3.2.2	语音信号的特点	91
3.2.3	语音信号的短时分析	92
3.3	语音识别的参量	94
3.3.1	语音信号的线性预测分析	94
3.3.2	线性预测倒谱系数	97
3.3.3	MFCC 系数	98
3.4	特定人小词表语音识别的动态规划算法	100
3.4.1	动态时间弯折算法原理	100
3.4.2	动态时间弯折的高效算法	103
3.5	非特定人语音识别的隐马尔柯夫算法	104
3.5.1	隐马尔柯夫过程应用原理	104
3.5.2	前向概率和后向概率——HMM 的输出概率计算	107
3.5.3	识别算法——Viterbi 解码	109
3.5.4	HMM 参量训练的 Baum-Welch 算法	110
3.5.5	多观察序列的训练算法	112
3.5.6	其他形式的 HMM	113
3.6	说话人自适应	114
3.6.1	说话人自适应概述	114
3.6.2	MAP 算法	115
3.6.3	MLLR 算法	117
3.7	大词表连续语音识别	120
3.7.1	搜索算法问题描述	120
3.7.2	动态规划搜索算法	121
3.7.3	剪枝操作	122
3.7.4	语言模型预判	123
3.7.5	基于词图的动态规划搜索算法	124
3.7.6	词对近似	125
3.8	说话人识别	126
3.8.1	说话人识别问题	126
3.8.2	说话人识别的方法	127
3.9	语音识别的其他算法	128
3.9.1	人工神经网络法	128
3.9.2	支持向量机法	135
3.9.3	差别子空间法	138
3.10	嵌入式系统中的语音识别	141

3.10.1	语音识别和嵌入式系统	141
3.10.2	算法的定点化	141
3.10.3	系统实现流程	142
3.11	应用系统实例	142
3.11.1	剑桥大学的语音识别工具包 HTK	142
3.11.2	卡内基·梅隆大学的语音识别软件包 Sphinx	144
3.11.3	五邑大学的噪声环境语音识别命令控制器	145
第4章	视觉——唇读与识别	147
4.1	唇读问题	147
4.1.1	唇读是语音的视觉表征	147
4.1.2	McGurk 效应	148
4.1.3	唇读感知系统的结构框图	148
4.2	图像的预处理	150
4.2.1	人脸图像主要特征位置的标定	150
4.2.2	人脸图像的跟踪	154
4.2.3	唇动定位和跟踪	155
4.3	唇动特征的提取	162
4.3.1	唇动特征的各种描述方法	163
4.3.2	函数可变模板灰度轮廓向量表征法	164
4.3.3	灰度轮廓权向量差分形状特征	173
4.4	唇读识别	174
4.4.1	视觉语音识别一般问题	174
4.4.2	DTW 法	176
4.4.3	HMM 法	179
4.4.4	TDNN 模型法	179
第5章	听觉-视觉——双模态语音识别与融合	180
5.1	双模态语音识别问题	180
5.2	双模态语音识别中的视觉语音特征区域定位	181
5.2.1	基于线性方法的视觉语音特征区域定位	181
5.2.2	基于支持向量机方法的视觉语音特征区域定位	186
5.2.3	基于核函数映射方法的视觉语音特征区域定位	188
5.3	视觉语音序列特征提取	193
5.3.1	变换处理	194
5.3.2	基于线性区别分析的特征参量投影	197
5.3.3	最大似然线性变换	197

5.4	基于隐马尔柯夫模型的双模态早期融合	200
5.5	基于隐马尔柯夫模型的双模态晚期融合	201
5.5.1	状态同步的双模态晚期融合中的 HMM	201
5.5.2	音节同步的双模态晚期融合中的 HMM	202
第6章	听觉-视觉——双模态语音识别数据库	205
6.1	多模态人机自然交互技术与数据库	205
6.2	双模态语音识别数据库的现状与发展前景	208
6.3	双模态数据库数据采集	211
6.3.1	数据库的语料设计与选择	211
6.3.2	数据库的采集	214
6.3.3	数据库原始数据的切分	216
6.4	双模态数据库管理系统设计	218
6.4.1	双模态数据库管理技术	219
6.4.2	双模态数据库的系统结构	220
6.4.3	双模态数据库的结构设计	223
6.4.4	数据库客户端应用程序设计	227
6.5	CAVBSR-WUIIS(1.0)数据库的使用与操作设计	229
6.5.1	CAVBSR-WUIIS(1.0)数据库系统的主界面及显示 方式设置	229
6.5.2	CAVBSR-WUIIS(1.0)数据库系统的各种查询	232
6.5.3	CAVBSR-WUIIS(1.0)数据库系统的记录添加	235
6.5.4	CAVBSR-WUIIS(1.0)数据库系统的记录的删除	237
6.6	双模态数据库在人机自然交互及身份认证中的应用	238
6.6.1	双模态数据库在唇读与人脸特征定位中的应用	239
6.6.2	双模态数据库在身份认证中的应用	241
6.6.3	CAVBSR-WUIIS(1.0)数据库的应用	243
6.7	数据库的扩展	245
	参考文献	246

Contents

Chapter 1 Introduction	1
1.1 Human-Machine Interaction and Human-Machine Nature Interaction	1
1.1.1 Human-Machine Interaction and Human-Machine Nature Interaction	1
1.1.2 Features and Functionalities of Human-Machine Nature Interaction	2
1.1.3 The Social, Science and Technology Background of Human-Machine Nature Interaction	3
1.2 The Interface of Natural Information Sources and Man Made Information	3
1.2.1 The Natural Information Sources	3
1.2.2 Man Made Information	4
1.2.3 Further Consideration	5
1.3 Interaction Object Recognition	6
1.3.1 Speaker Recognition	6
1.3.2 Approaches to Speaker Recognition	7
1.4 Interaction Content Recognition	11
1.4.1 The Key Problems in Interaction are the Interaction Content Recognition	11
1.4.2 Natural Language Understanding	15
1.5 Audio-Visual Bimodal Fusion	16
1.5.1 Fusion Problems	16
1.5.2 Fusion Strategies	16
1.5.3 Fusion Strategies and Recognition Algorithms	17
1.6 Apperception of Human Emotion States	18
1.6.1 Audio-Visual Bimodal Emotion State Recognition Problems	18
1.6.2 The Explicit Channel and the Implicit Channel	19

1.7	Multimodal Network, Localization, Tracking and Database	20
1.7.1	Human-Machine Nature Interaction System	20
1.7.2	Network Environment	21
1.7.3	Speaker Localization and Tracking	22
1.7.4	Bimodal Database	23
1.8	Changes of Production Methods, Working Styles and Life Styles of Human Beings Brought by Human-Machine Nature Interaction	25
1.8.1	The Implementation of Human-Machine Nature Interaction Will Bring Changes	25
1.8.2	Potential Military Application and Civil Application	25
Chapter 2	Visual—Speaker Recognition and Face Recognition	27
2.1	Speaker Recognition Problem	27
2.2	Preprocessing of Human Face Image	29
2.2.1	Detection and Localization of Human Face Image	30
2.2.2	Standardization of Human Face Image	30
2.3	Feature Extraction and Recognition of Human Face Image	37
2.3.1	Geometry Feature Method	38
2.3.2	Fisher Face Method and Local Feature Method	39
2.3.3	Elastic Model Method	39
2.3.4	Neural Network Method	40
2.3.5	Invariant Moment Feature Method	40
2.4	Adaptive Principal Components Extraction Algorithm (APEX)	42
2.4.1	Statistical Principal Components Analysis	43
2.4.2	APEX	43
2.4.3	Convergent Analysis of APEX	45
2.4.4	Application Examples	49
2.5	Face Image Singular Value Feature Extraction Method	55
2.5.1	Singular Value Feature	55
2.5.2	Singular Value Dimension Reduction	56
2.5.3	Application Examples	57
2.6	Optimal Discrimination Vector Feature Extraction	63
2.6.1	Kernel Function Fisher Discrimination	63
2.6.2	Generalized Kernel Function Fisher Optimal	

Discrimination	66
2.7 Image Segmentation Algorithm in Face Recognition and Application Examples	72
2.7.1 Human Face Image Segmentation Algorithm for Face Recognition	73
2.7.2 Human Face Image Fast Segmentation Algorithm for Face Recognition	80
Chapter 3 Audio—Speaker Recognition, Speech Recognition and Understanding	86
3.1 Speech Recognition Problems	86
3.1.1 Speech Recognition Technology Development	86
3.1.2 Classification of Speech Recognition System	89
3.1.3 Structure of Speech Recognition System	89
3.2 Speech Signal Features	90
3.2.1 Digitalizing of Speech Signal	90
3.2.2 Speech Signal Features	91
3.2.3 Short-Time Analysis of Speech Signal	92
3.3 Speech Recognition Parameters	94
3.3.1 Linear Prediction Analysis of Speech Signal	94
3.3.2 Linear Prediction Cepstrum Coefficient	97
3.3.3 Mel-Frequency Cepstrum Coefficient	98
3.4 Dynamic Programming for Speaker Dependent Recognition	100
3.4.1 Principles of Dynamic Time Warping	100
3.4.2 Fast Implementation of Dynamic Time Warping	103
3.5 Hidden Markov Model for Speaker Independent Recognition	104
3.5.1 Principles of Hidden Markov Model	104
3.5.2 Evaluation of Output Probability	107
3.5.3 Viterbi Decoding for Recognition	109
3.5.4 Baum-Welch for Training	110
3.5.5 Multi-Observation Training	112
3.5.6 Other Hidden Markov Models	113
3.6 Speaker Adaptation	114
3.6.1 Overview of Speaker Adaptation	114
3.6.2 Maximum a Priori Adaptation	115
3.6.3 Maximum Likelihood Linear Regression Adaptation	117