新一代(第四代)人机交互的概念框架特征及关键技术

袁保宗,阮秋琦,王延江,刘汝杰,唐晓芳
(北京交通大学信息科学研究所,北京 100044)

摘 要: 人机交互是研究人与计算机之间通过相互理解的交流与通信,在最大程度上为人们完成信息管理,服务和处理等功能的一门技术科学.本文给出新一代(第四代)人机交互的概念框架和其四个标志性特征,即:(1)具有多模感知(听觉,视觉,手势,笔势等)功能的人机交互方式;(2)可进行基于 Agent 的听、视觉对话,作为人机交互的界面;(3)具有 Internet 数据仓库和基于内容检索的知识处理能力,作为人机交互内容;(4)可以在二维或虚拟的三维环境中实现人机通信,作为人机交互的环境.文中给出了解决和实现该四大特征的关键技术,描述了实验系统平台的全面功能,可以为同类研究人员作参考.

关键词: 人机交互; 多模感知; 多功能感知; 智能 Agent; 口语对话; 信息融合计算
中图分类号: TP302.7 **文献标识码:** A **文章编号:** 0372-2112 (2003) 12A-1945-010

A Conceptual Model and Features of New Generation (Fourth Generation) Human-Computer Interactive Systems

YUAN Bao-zong, RUAN Qiu-qi, WANG Yan-jaing, LIU Ru-jia, TANG Xiao-fang
(*Institute of Information Science, Northern Jaiotong University, Beijing* 100044, *China*)

Abstract: Human-computer interaction is a complex of technical science, which investigates the interaction and communications through mutual understanding between human and computers to solve problems of information managements, services and processing in maximized degree for human needs. This paper proposed a conceptual model of new generation (fourth generation) human-computer interaction systems and their four major features: (1) the ways for interaction are through multimodel perceptions(hearing, vision, gestures and pen writing); (2) the interface for interaction is through the intellegent agents, who may understand user's requirements and implement all required processing by machine through auditory/visual diologues; (3) the contents for interaction may be performed by techniques of content based text or image/video retrieval; (4) the display environment of interaction may be realized in 2-D or 3D through virtual browsing the 3-D scene of computer graphics reconstructed by 3-D computer vision technique. In this paper the possible solutions and algorithms to implement these four major features are given and the platform of the 4th human computer interative system is also described.

Key words: human-computer interaction; multimodel perceptions; multifunction perceptions; intellegent agents; spoken language dialogue; information fusion calculation

1 引言

人机交互(Human-computer interaction,简称 HCI)是研究人与计算机之间通过相互理解的交流与通信,在最大程度上为人们完成信息管理,服务和处理等功能,使计算机真正成为人们工作学习的和谐助手的一门技术科学,它是伴着计算机的诞生就发展起来的.在现代和未来的社会里,只要有人利用通信、计算机等信息处理技术,为社会、经济、环境和资源进行活动时,人机交互都是永恒的主题.鉴于它对科技发展的重要性,研究如何实现自然、便利和无所不在的人机交互,成是现代信息技术,人工智能技术研究的至高目标,也是数学,信息科学,智能科学,神经科学,以及生理、心理科学多科学交叉的新结合点,并将引导着二十一世纪前期信息和计算机研究的热门方向.

人机交互技术的发展与国民经济发展有着直接的联系,它是使信息技术融入社会,深入群体,达到广泛应用的技术门槛.任何一种新交互技术的诞生,都会带来其新的应用人群,新的应用领域,带来巨大的社会经济效益.这一点从移动通信

收稿日期:2003-09-18;修回日期:2003-12-10
基金项目:国家自然科学基金重大项目(No.9690015);重点项目(No.69789301);国家自然科学基金面上项目(No.69775003)

中手机的发展和多媒体电脑给社会带来的效益就可以看到.未来新一代具有自然人机交互功能的随身电脑,将成为所有专业和普通人群生活,学习,工作的必备助手和工具,并对改变社会生活与工作的模式起到深远的影响.正是这个原因,世界各国信息领域的学者和企业巨头都部署了发展新一代人机交互领域的中长期发展计划.

在人类科技发展历史中,任何一门技术都存在着不同的发展过程.随着生产力的发展及人类对大自然的认识的深入,每一种技术都将不断地提高和前进.为了规范某一种技术的发展,国际上相应的组织往往会对该技术的发展阶段进行相应分代的划分,以便制定标准,统一研究方向.人机交互技术涉及到广泛的计算机应用,面临国际化,也需要有一个国际化的、统一的概念框架,以便为进一步的制定标准做准备.因此,对人机交互技术的各个发展阶段进行相应的代别划分同样有着积极的促进意义.

人机交互作为计算机科学研究领域中一个重要组成部分,其发展历程已经经历了半个多世纪,并且取得了很大的进步和提高.我们认为,从计算机的诞生之日起,人机交互技术的发展已经经历了三个阶段:

(1)基于键盘和字符显示器的交互阶段 这一阶段所使用的主要交互工具为键盘及字符显示器,交互的内容主要有字符、文本和命令,交互过程显得呆板和单调.这一阶段可称为第一代人机交互技术.

(2)基于鼠标和图形显示器的交互阶段 这一阶段所使用的主要交互工具为鼠标及图形显示器,交互的内容主要有字符、图形和图像.70 年代发明的鼠标,极大地改善了人机之间的交互方式,在窗口系统大量使用的今天几乎是必不可少的输入设备;应该说,鼠标和窗口系统的出现,是人机交互技术发展历史上的一次技术革命.这一阶段可称为第二代人机交互技术.

(3)基于多媒体技术的交互阶段 20 世纪 80 年代末出现的多媒体技术,使计算机产业出现了前所未有的繁荣,声卡、图像卡等硬件设备的出现使得计算机处理声音及视频图像成为可能,从而使人机交互技术开始向声音、视频过渡.在这一阶段,人机交互的工具除了键盘和鼠标外,话筒、摄像机及喇叭等多媒体输入输出设备,也逐渐为人机交互所用;而人机交互的内容也变得更加丰富,特别是语音信号处理技术的发展,使得通过声音与计算机进行交互成为可能.多媒体技术使用户能以声、像、图、文等多种媒体信息与计算机进行信息交流,从而方便了计算机的使用,扩大了计算机的应用范围.另外,多媒体技术的发展,促进了信息处理技术特别是计算机听觉与计算机视觉的发展,从而使人机交互在朝着自然、和谐的方向上向前迈进了一大步.多媒体交互阶段可称为第三代人机交互技术.

(4)基于多模态技术的交互阶段 在第三代人机交互技术中,多媒体技术的发展虽然提供了多媒体信息处理的可能性,但是就当前发展的现状,仍处于独立媒体的存取;编辑及媒体间的并合水平,尚未涉及多媒体信息的综合处理.实际上从单媒体走向多媒体,它可能发生的作用决不应该是它们间的量的简单合并,而应该是一种质的根本飞跃.虽然通过多媒体信息进行人机交互极大地丰富了人机交互的手段和内容,但离人类之间的交互能力还差得较远.我们知道,人类在与其环境进行交互时是多模态的.人可以同时说、指和看同一个物体;还可以通过同时听一个人的说话语气和看他的面部表情及手臂动作来判断他的情绪.为了更好地理解周围的环境,人类每时都在使用视觉、听觉、触觉和嗅觉,可以说多模态是人类之间自然交互的体现.此外,人类之间的交互是基于知识的,并主要是通过语言进行交流.因此,基于多模态技术的交互阶段将是第四代人机交互技术.这正是本文所要研究的主题.

早期的多模态人机交互系统之一是由 Bolt[1] 提出来的,他的"Put-That-There" 系统融合了话音输入和三维手势两种人机交互模式.这一系统的建立为多模态人机交互的后续研究提供了范例.如 M Fukomoto 等研制的 Finger-Pointer[2] 同样也集成了三维手势和语音命令用于控制视频的展示.

QuickSet[3,4] 是一用于军事仿真控制的多模态人机接口系统.它采用语音和基于笔输入的手势符号作为人机交互的模式,用户可以同时用笔输入和语音创建军事实体、建立控制措施、标记控制地区等.该分布式仿真系统是为美国海军陆战队研制的,并已经得到应用.多模态地图(Multimodal maps)[5] 是由斯坦福大学人工智能研究中心的 A Cheyer 等人为研究多模态人机接口技术而构造的一个旅行计划安排系统.用户可同时使用语音、手势符号、手写文本与系统进行交互,而系统的输出则可通过视频、文本及声音呈现.由于与地图进行交互容易集成语音、笔势、文本等多种交互模式,因此近几年来,类似的研究很多,如 CMU 大学系统交互实验室的 A Waibel 教授等[6,7],以及美国 Oregon 研究生学院 S Oviatt 教授等[8,9] 都对此类系统进行过深入的研究.另外,近年来基于语音和唇读的多模态人机交互系统的研究也受到了普遍的重视,其主要目的是提高在嘈杂环境下的语音识别率.如 A Adjoudani 等[10] 分别对基于 HMM 的特征级融合和基于最大似然估计的决策级融合方法用于语音识别进行了研究.

近年来随着计算机视觉与图像处理技术的发展,基于视觉的眼睛及视点跟踪技术日益受到重视,并被引入到多模人机交互之中.如 M Akay 等[11] 构造的一个多模态人机交互系统集成了语音、手势符号及眼睛跟踪等多种交互方式用于远程医疗分析和决策.S Pastoor 等[12] 构造了一个三维数据可视化多媒体实验系统,该系统可通过眼睛注视的方式控制所要查看的信息.K Tsui 等[13] 构造了一个基于多模态信息的工作调度管理系统,该系统可以输出声音、图形和文本等信息.另外,N Tosa 等[14] 研制了一个多模态自主智能 Agent - NeuroBaby,该 Agent 可以识别人的语气和声调、跟踪人的眼睛和手势,并能够通过表情和声音与用户交流.该系统目前已经可以通过网络实现不同文化背景的两个用户之间的非语言交流.在计算机视觉和三维可视化方面,如 MIT 媒体实验室以 A.Pentland[15] 为首的智能感知(Perceptual intelligence)研究组所从事的智能屋(Smart rooms)的研究.人在此屋内可以用语言、表情或手势与计算机进行交互,如控制运行计算机程序,浏览多媒

体信息或进入虚拟环境与虚拟生物进行交互等.除了 Smart Rooms 的研究外,Pentland 还领导研究了类似的智能工程,如智能椅子、智能服装、智能汽车等.MIT 的媒体实验室,也开始了类似的研究.CMU 交互系统实验室(Interactive system lab)在 A Waibel 教授的领导下,开展了诸如计算机视觉、人脸检测跟踪、注视跟踪、手势识别等多项智能人机交互技术的研究[16].另外像 CMU 计算机科学系的 VASC 三维形状合成项目[17],马里兰大学人机通信接口实验室(HCIL)的虚拟环境系统[18],瑞士日内瓦大学的 MaraLab 的人物合成[19,20],以及英国爱丁堡大学电机系通信接口研究中心(CCIR)的信息检索系统等[21],在图形方面均做出了很高水平的工作,但是这些工作缺少与听、视觉信息交互的配合.

在国内也有相应的研究,如高文教授等[22,23]提出了一个基于多模态接口技术的聋人与正常人交流系统,从而使聋哑人能够与正常人进行交流.

综上所述,目前多模态人机交互技术的研究主要体现在以不同应用出发而开发的不同系统,缺少一个统一的、能够代表人机交互技术较为完整、具有指导意义的基本框架.为此,本文结合我们目前在计算机视觉、计算机听觉、虚拟现实、模式识别及人工智能等领域进行的研究,对这个问题进行了研究和探讨,并提出了第四代人机自然交互与通信的概念框架及其特征描述.

2 第四代人机自然交互与通信的概念框架及其标志性特征

作为新一代人机自然交互系统应该具备以下四种功能:(1)多模感知功能;(2)智能代理交互功能;(3)知识处理功能;(4)可视化显示功能.这四种功能也正是人类之间交互的具体体现.我们知道人类之间的交互是多模态的,人类每时每刻都在使用视觉、听觉、触觉和嗅觉,可以说多模态是人类之间自然交互的体现.此外,人类之间的交互是基于知识的,并主要是通过语言进行交流.实际上,在人机自然交互中智能代理的作用就是要模拟用户的交互对象,使其像人一样具有多模感知和知识处理能力,并能以对话的形式与人进行交流.而可视化功能,主要是为了体现人机交互的真实感和沉浸感.如果没有真实场景的可视化显示,人机交互过程将显得枯燥和乏味,可以说这四个方面是缺一不可的.

据此,我们提出了新一代——第四代人机自然交互与通信的概念.它是一种以多模信息交互为输入输出,以 Agent 为交互通信界面,具有基于知识对话的网络信息交互和检索能力,并具有二维和三维虚拟交互环境可视化显示的人机交互技术.

它的概念模型可用图 1 表示.

该概念模型主要包括以下几个部分:

(1)多模态输入/输出　多模态输入/输出是第四代人机交互与通信的主要标志之一.多模态输入包括键盘、鼠标、文字、语音、手势、表情、注视等多种输入方式;而多模态输出包括文字、图形、语音、手势、表情等多种交互信息.多模态输入/输出极大地丰富了人机交互的交互方式,使得人机交互更加

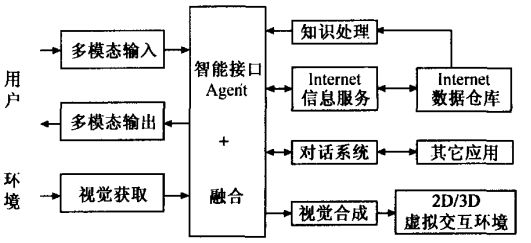

图 1　第四代人机交互通讯概念模型

自然、简单,更加趋向于人类之间的交互.

(2)智能接口代理　智能接口代理是实现人与计算机交互的媒介,用户的交互任务均可由它来完成.理想的智能接口代理是一个类人化的软件实体,可由三维人脸模型标征,具有唇动和表情功能,且能够感知并理解用户的交互信息,并对用户的交互要求作出反应.

(3)视觉获取　人机自然交互有时需要对人机交互环境进行实时重构,这就要求计算机系统能具备视觉功能,特别是主动视觉功能.计算机视觉是计算机科学和人工智能的一个重要分支,代表了人类视觉理论和计算机理论两方面成果的结晶.而主动视觉是对人类视觉更高认识的基础上提出和实现的新方向,它更接近人类视觉系统.视觉系统主要用于实时获取外部视觉信息.

(4)视觉合成　在第四代人机交互与通信系统中,需要将现实中的场景,如实验室、会议室、银行、机场等在计算机系统中真实地再现出来,这就需要系统应具有视觉合成功能.具有从真实现实到虚拟现实的转换能力(From actual reality to virtual reality,简称 AVR),从而使人机交互能够在一个仿真或虚拟的环境中进行,仿佛现实世界中人与人之间的交互.

(5)对话系统　口语对话是人与人之间进行交流的最主要和最自然的方式.对话系统的研究目前主要有两种趋势,一种是以语音为主,结合其他交互手段如手势、表情、唇读等;另一种是从某一特定任务域入手,引入对话管理概念,建立类似于人人对话的人机对话,研究对话系统的通用性和可移植性.对话系统涉及到语音识别、句法处理、知识表示、语句生成、语音合成及对话建模等许多关键技术.

(6)Internet 信息服务　随着计算机网络的迅速发展和普及,通过互联网进行信息服务和信息交流已经成为信息时代的一个重要的标志.传统的人机交互观念已经不能适应,而基于网络化的、分布式的以人为中心的人机交互与通信系统将成为研究的主导方向,计算机在这样一个系统中将扮演信息交流媒介的角色.

(7)知识处理　Internet 的迅速扩展延伸,大量的无组织的数据和海量的信息常常使用户无从下手,难以立即找到自己所需要的信息.在这种情况下,就要求能通过一定的技术手段,自动地提取有组织的、可为人们利用的知识,因此如何对知识进行处理就成为新一代人机交互与通信技术中的一个非常关键的问题.

综上所述,我们可以将第四代人机自然交互与通信的主要内容概括为以下四个标志性特征:

(1)具有多模感知(听觉,视觉,手势,笔势等)功能,完成人机交互方式;

(2)可进行基于 Agent 的听、视觉对话,作为人机交互的界面;

(3)具有 Internet 数据仓库和基于内容检索的知识处理能力,作为人机交互内容;

(4)可以在二维或虚拟的三维环境中实现人机通信,作为人机交互的环境;

我们按人机交互手段和交互内容为主线,把以多功能感知机为手段的交互系统提高到第四代人机自然交互与通信的概念上加以认识,明确地提出了它的概念内涵,并将其与前三代人机交互技术进行了比较,如表 1 所示.从中可以看出,与前三代人机交互技术相比,第四代人机交互与通信技术无论是交互手段还是交互内容,都产生了质的飞跃.在第四代人机自然交互与通信的层次上,人机交互过程将更加自然化、智能化.

表 1 四代人机交互与通信技术比较

	交 互 手 段	交 互 内 容
第一代	键盘,字符显示屏幕	文字,数字,符号,键控命令
第二代	鼠标,图形显示屏幕	字符,图形,图像
第三代	多媒体输入(话筒,摄像头),喇叭,HMD	声音,图像,话音
第四代	多模态感知输入,Agent,3D 虚拟环境,自动知识获取	自然对话,手势表情,个性化知识检索服务

上述是第四代人机自然交互通信的全部内涵.可以看出,第四代人机自然交互与通信的概念、模型的建立是综合了现代信息科学和模式识别人工智能科技成果,它是一个开放式的平台环境,允许不断集成最新研究成果,以期提高人机自然交互的和谐程度.

3 若干关键技术的解决和实现

3.1 人机自然交互多 Agent 结构

第四代人机自然交互涉及到多模态信息输入、多模态信息融合、多模态信息输出、视觉信息获取与理解、视觉信息重建与显示、大规模数据库、知识处理、用户智能代理及各种应用服务.可见系统是由许多独立的功能部件所组成,每一个部分在独立完一定功能的同时,还需要实时与系统中其它部分进行通信与合作,以实现某些自己所不能独立完成的功能,完成整个交互过程和任务处理的关键是通信与合作.针对这样一个强调分布、合作与通信特点的系统而言,单纯依靠传统的软件开发和系统集成理论和技术进行其研制开发工作是很难达到满意的效果,而 Agent 则在这些关键问题上显示其优越性.利用 Agent 的自主性、可交互性和协作性甚至是学习性可以简化这样一个大型系统的研制与开发.

为此,我们可以把第四代人机自然交互中的每一个独立部分都看作是一个特定的 Agent,整个系统本身则是由许多协作型 Agent 所组成,系统整体功能的实现依赖于各个 Agent 之间的交互与协作.具体地,基于 Agent 的第四代人机自然交互与通信的结构模型如图 2 所示.

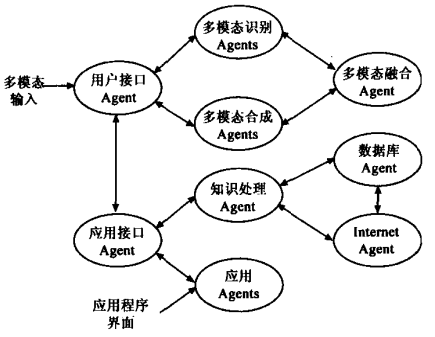

图 2 第四代人机自然交互与通信多 Agent 结构模型

其中,多模态识别、多模态融合及多模态合成部分的详细结构如图 3 所示.

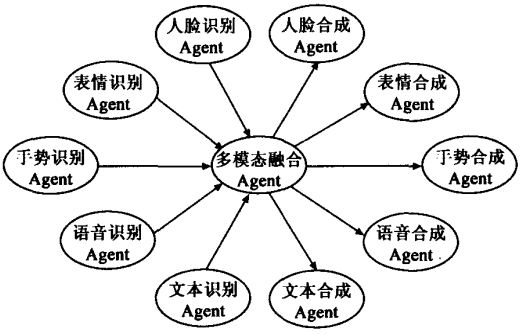

图 3 多模态感知部分 Agent 结构模型

图中箭头指示出了 Agent 间的依赖和合作关系.如用户接口 Agent 负责处理多模态输入信息及多模态输出信息,并与各多模态识别 Agent 和多模态合成 Agent 建立通信联系,以便对多模态信息进行识别、理解与合成,从而达到能与用户进行基于视听信息的交互.此外它还负责与应用程序接口 Agent 的通信与交互,以便根据不同的应用选择不同的程序界面以及不同的数据库和知识库.Agent 之间在能够进行通信以前必须建立一套标准的通信原语,包括语法结构及命令格式.在上述人机自然交互系统结构中,我们基于 FIPA KQML[24],构造了一种简化的多模态 Agent 通信语言 —— MMACL .基于上述通信语言及机制,我们使 Agent 具备:(1)Agent 被赋予承担一些任务和具有求解某些问题的能力,并能决策如何及时完成任务和提交结果;(2)Agent 能够与其他 Agent 在求解环境中共存,并以某种方式相互作用;(3)Agent 具备所在求解领域中的知识,以及求解环境中其它 Agent 的一些信息, Agent 的基本构造模型[25],如图 4 所示.

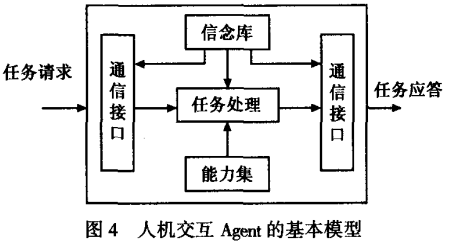

图 4 人机交互 Agent 的基本模型

在该模型中,Agent 由通信接口、能力集、信念库及任务处理四个部分构成.其中,通信接口用于 Agent 之间进行交互与联络,如处理任务请求或应答;能力集表示 Agent 能干什么,并提供相应的处理过程和算法;信念库用于表示 Agent 的世界模型,即 Agent 对其环境所拥有的知识,也称本体知识(ontology);而任务处理则根据 Agent 的能力利用相应的处理算法及信念知识完成所请求的任务.下面是一个基于语音识别 Agent 和手势符号识别 Agent 的多模态主动视觉控制的例子,当用户通过话筒输入声音信息时,语音识别 Agent 会通过相应的消息响应函数,自动获取语音信息,然后调用相应的识别算法进行识别,在完成识别后,返回相应的文本.语音识别 Agent 的运行界面如图 5 所示.

手势符号识别 Agent 的运行界面如图 6 所示.

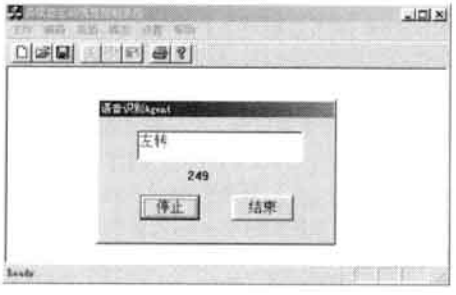

图 5　语音识别 Agent 界面

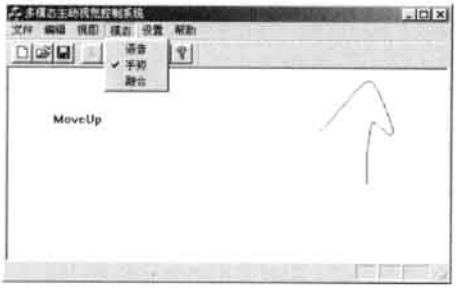

图 6　手势符号识别 Agent 界面

3.2　人脸检测,跟踪与识别

第四代人机自然交互主要体现在用户和用户智能代理之间的交互,用户的任务可以由其代理负责处理.任务完成过程中,用户需与其代理不断地交互.因此,用户代理除了能识别用户的语音信息外,还应能识别用户的脸形及面部表情,从而更好地理解用户的意图.所以,有关人脸面部感知机理的研究也是第四代人机自然交互与通信中的一个关键技术之一.而要完成对人脸的感知,首先要解决的一个问题是人脸检测与跟踪,特别是复杂背景下的人脸检测与跟踪问题的研究.因为人机交互的环境一般是复杂的和随意的,因此对如何快速而又鲁棒地在任意的环境中正确检测到人的面部的研究是极其重要的.

本文在肤色模型参数的基础上提出了一种基于进化 Agent 计算的区域分割和标记方法[26,27].进化计算是指以达尔文进化原理为仿真依据,在计算机上实现的具有进化机制的算法和程序,它的主要研究内容包括遗传算法、进化策略、遗传规划、进化编程及分类系统等.所谓进化 Agent 是指 Agent 能够适应环境并在环境内实施进化行为,具体地,它主要具有以下功能:(1)它可以自动感知其在环境的状态,获知自己在环境中的位置;(2)它可以根据所感知的信息,判断自己所处的位置是否属于问题的解空间,如果是,它将记录下该位置,并在此位置附近实施进化行为;(3)它具有有限的生命期,当它的生命超过规定的期限时,它将在环境中消失.因此一个进化 Agent 也可用一八元组来表示:

$$Agent = \langle p, d, a, f, fml, Diff, Rep, Die \rangle$$

其中 p 代表它在环境中的位置,d 表示它在环境中当前的移动方向,a 代表其年龄,f 表示它的适应值,该值表示该 Agent 的优劣程度,*fml* 表示 Agent 的家族标志.以上这些参数表示 Agent 的内在状态,而 *Diff* 和 *Rep* 分别代表 Agent 的扩散和复制进化行为,*Die* 表示一个 Agent 象生物一样具有一定的生命期,它能够从环境中消失(死亡).图 7 给出了进化 Agent 的计算流程.

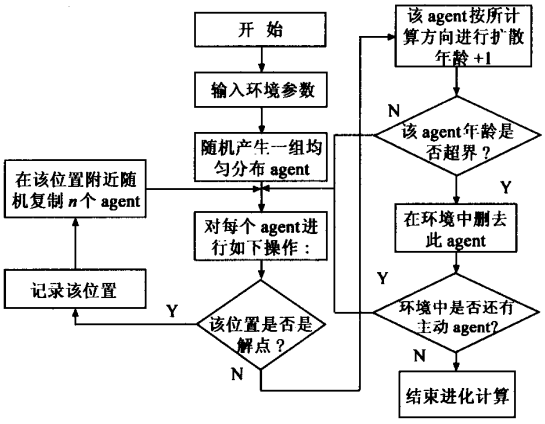

图 7　进化 Agent 的计算流程

为确定人脸区域,使用了一个简单的正面人脸模型,利用该模型及其各种变形可以很容易地定位出具有一定姿态变化的人眼;然后通过如何利用小波变换得到人脸特征的分布和基于神经网络的人眼定位方法[28,29].图 8 是部分人脸检测结果,图 9 是复杂情况下的人脸检测结果.

我们利用遗传算法对人脸进行跟踪,跟踪效果见图 10.

经过人脸检测与跟踪后,系统进入人脸识别和信息融合,这部分见后续信息融合小节.

3.3　人机口语对话[30~33]

人机口语对话不同与一般的口语查询,它涉及自然语言理解,对话模型和对话管理等技术,如图 11 所示.

从图中我们可以看到,人机交互过程大致可以分为四层进行.第一层是包括语音识别、理解和文语转换(语音合成)的语音信号处理层.自然语言处理包括自然语言理解和自然语言生成,构成人机交互的第二层.计算机理解自然语言,并与用户进行对话是交互过程的第三层即对话处理.通过对对话处理,可以使计算机能处理交谈过程中的语法、语义、语气等方面的因素,准确的理解用户的输入,掌握其意图,有效的生成满足交谈语境的话语,并控制交谈的过程.对话处理属于人

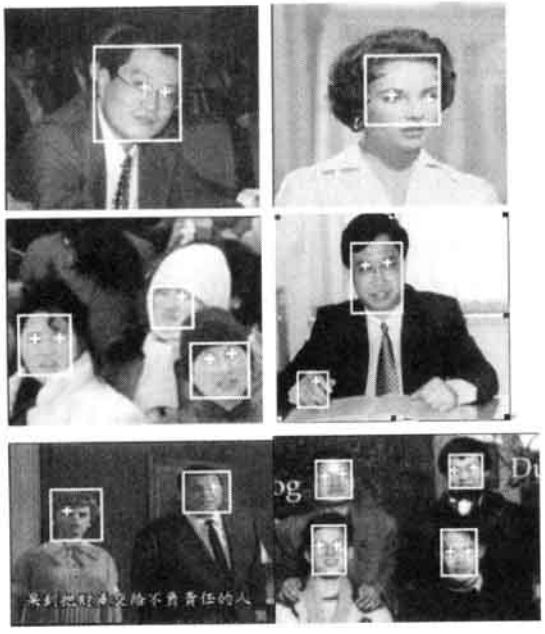

图 8 部分人脸检测结果

图 9 复杂情况下人脸检测结果

图 10 基于主动进化 Agent 的人脸跟踪

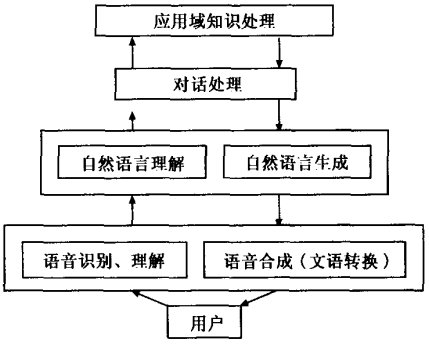

图 11 人机自然交互过程的分层结构

机交互的核心位置，它架起了输入与输出之间的桥梁.目前，人机交互中最迫切需要实现的对话是面向任务的对话，它是指围绕某一主题(即任务或目标)，有明确目的并为了实现这一目的而进行的对话.第四层：应用域知识处理，它是处理应用域知识所必需的.

面向任务的对话处理包括了任务域的分析和对话规律的研究.协调对话进程是对话处理特有的内容，任务域的分析是解决协调问题的基础，它要处理好如何实现任务和预测用户需求两大问题.对话系统结构有对话元素的建立，对话优先权(主动权)控制，系统的上下文反馈，用户关注焦点预测，对话记录，用户模型建立，即对用户知识的动态描述，和辅助性的对话的插入等.图 12 是一个具有自然语言接口的智能多媒体问询应答系统(IMIAS)，所选择的任务域为公园导游域的框图.

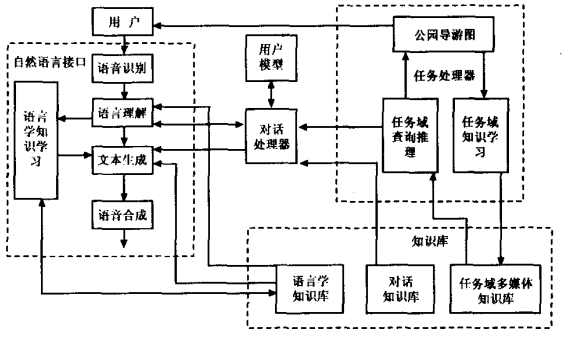

图 12 公园导游域的口语对话系统

上图中的对话处理器功能有：(1)接受理解模块提供的用户话语语义的形式化表述；(2)进行语句类分析，得到用户意图；(3)激活任务处理器，向其提供任务要求；(4)选择子对话；(5)确定优先权级别；(6)向生成模块提供应输出的语义的形式化描述；(7)预测用户下一个输入的语义；(8)更新用户模型.任务处理器完成：(1)学习任务域知识，包括公园导游图的填充，公园及其中景物的特征知识，有关的多媒体知识等.学习结果存入任务域多媒体知识库；(2)接受对话处理器的任务请求，完成任务推理，并产生预测；(3)查询任务域多媒体知识库，向对化处理器或用户产生输出.自然语言接口根据对话处理器提供的对用户输入的预测，识别用户的话语，理解其含义，并向对话处理器提供用户语义的形式化表述，接收对话处

理器提供的需要输出的语义,生成汉语文本输出,或语音合成模块转换为语声输出,完成对语言学知识的学习,既当系统无法"理解"用户的话语时,对其句式及其中的词汇进行学习,将学习结果存入语言学知识库.知识库包括 任务域多媒体知识库,公园及公园内景物的位置和特征;与公园及各景物相关的多媒体材料;用户模型:(1)用户已知信息;(2)用户关心的地点;(3)用户关心的特征;(4)用户意愿;(5)用户当前位置;(6)用户动作.系统性能评价可以达到:(1)允许多方式的自然语句输入,对同一问题多问多答,对任何一种问题, 可随即生成 2~15 种不同的输出;(2)包含词汇一千多;(3)覆盖汉语中大多数单句和复句;(4)具有学习功能;(5)具有自组织能力;(6)知识库便于添加、易于维护.基于以上原理的颐和园导游口语对话系统已在实验室中运行多年,性能良好.

3.3 基于内容的信息检索[34~36]

文本检索为人们查找信息、获取知识提供了便利,目前 Web 上用于检索的主要工具是搜索引擎,其核心技术已被广泛应用.但搜索引擎有明显的弊端,搜索结果庞大,内容不准确等.于是更高层次的检索,如文本挖掘成为当今的研究热点.在基于内容的文本信息检索中首先要对互联网的文本进行分类.支持向量机在用于分类任务时有良好的性能.我们提出了两种改进的支持向量机方法(NN-SVM,LSVM),在复旦大学李荣陆提供的小规模中文文本语料(中文自然语言处理开放平台,http://www.nlp.org.cn/)上进行了分类实验,语料经过分词、词频统计、特征提取等处理后,共选取了 1221 个特征词,分类正确率达到 96.3%.为了提高利用语音进行检索的正确性,还提出了一种新的文本串文音相似度的定义,大大提高了基于语音的文本检索正确率[37,38].

图像检索也是基于内容的信息检索的重要部分.我们提出了两种算法,即基于彩色距的非检督多分辨率图像分割检索算法和基于粗糙集的图像检索算法.前者利用搜索显著块(Salient block) 将图像的前景与背景分离,再用多尺度分析进行分类搜索,而后者是通过对图像视觉特征和语义特征的基于粗糙集约简知识推理,以规则分类进行检索.图 13 给出了检索的结果.

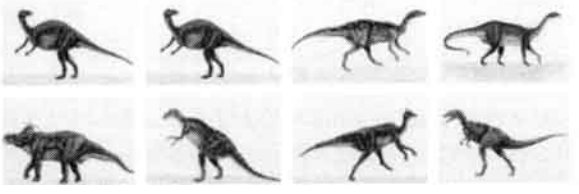

图 13　检索结果:第一幅为输入样本其它为检索输出

3.4 三维虚拟环境的建立[39~42]

为建立三维虚拟环境,需要解决从数字图像或视频拍摄设备,利用计算机通过任意二维序列未定标图像获得三维景物信息,进而进行三维重建,最后实现三维场景的虚拟浏览,以往的三维信息获取侧重于研究定标系统下物体的三维重建,由于静态定标方法使用的复杂性,本文以研究了未定标系统下序列图像运动和结构恢复问题,采用对极几何约束,三线性张量关系,以及基于不同参照系的摄像机定标技术,应用于物体的三维重建和建模,和主动式表面重建,视点合成等众多方法和技术,实现了一个从三维空间真实环境到三维虚拟环境实时浏览的系统,称为 AVR 系统 (AVR 是 From actual to virtual reality 的缩写).

AVR 重建算法的基本流程如图 14.

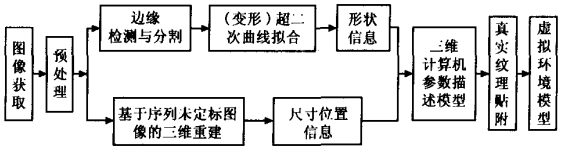

图 14　AVR 重建算法的基本流程

我们这里着重强调的是 AVR 理论中最具特色的内容,即超二次曲线、超二次曲面以及三维计算机参数描述模型.我们若仔细观察现实世界,不难发现,现实世界的物体大都可以考虑成简单的几何体或简单几何体的组合.对于物体的轮廓往往可以用一条闭合曲线——超二次曲线——进行拟合描述;对于整个物体表面可以选择一合适的闭合曲面——超二次曲面——加以表达.而对于整个场景,可以使用基于超二次曲面的层次式描述.由于超二次曲面强大的描述能力,极其适合于做物体的参数化表示,同时也可以用作物体识别.为了增强超二次曲面的几何表现能力,通常还考虑沿 Y 轴的锥化变形与沿 Y 轴弯曲变形.超二次曲面仅用 13 个参数就可以表示一个物体,因此,参数数据非常稀少.对于表面形状复杂、曲率变化丰富的物体无法进行准确的描述,我们提出了基于超二次曲面的层次化描述,包括四个描述层次:即采用单个超二次基元的描述;采用超二次基元 CSG 布尔组合的描述;使用全局 DMFFD(Direct manipulation of freeform deformation)对超二次基元描述模型进行全局调整;在表面形状曲率变化复杂的区域进行局部 DMFFD 调整.这样就可以对真实世界景物进行不同层次,不同细节程度的描述.它具有较多优点:例如(a)使用参数化模型对虚拟环境建模,数据量小,描述精练;(b)使用超二次曲面作为虚拟环境建模的基元模型,超二次曲面模型表现力很强,通过基元的 CSG 布尔组合,可只使用少数几个参数,描述大量的真实世界中具有相对规则形状的人工景物,而这些相对规则景物构成了真实世界景物的主体;(c)使用直接控制自由变形 DMFFD 对超二次曲面基元描述模型进行全局和局部精化调整,可实现对复杂形状物体表面曲率变化的描述;(d)实验层次描述,可实现对物体从粗略到精细的描述,符合虚拟环境建模的多细节层次(LOD:Level of detail)描述方式,因此,满足了虚拟环境的高层处理的要求.图 15 与图 16

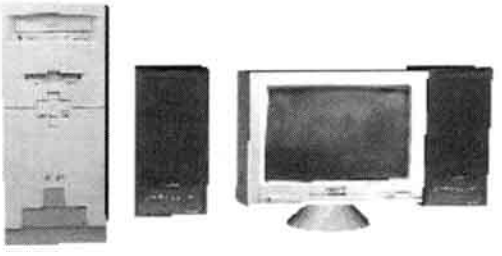

图 15　重建后的真实场景在浏览器中的显示

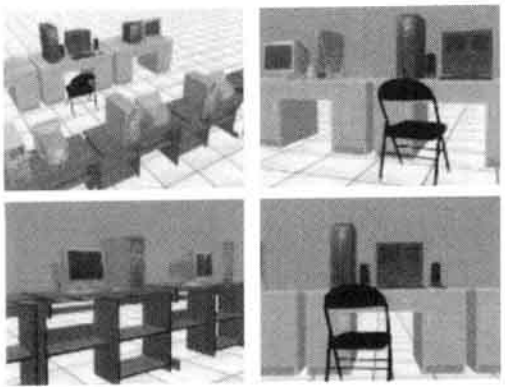

图 16 实验室环境 AVR 重构及漫游

是利用 AVR 重建的实物计算机和实验室.

3.5 多感知信息融合计算

作为第四代人机交互的核心内容之一是多感知信息融合计算模型.为使机器真正具有自然交互的要求,系统必需有一定的认知和记忆能力,多感知信息融合的计算模型就应该是这种能力的体现.图 17 给出了人们大脑中信息融合模型示意图[43,44].

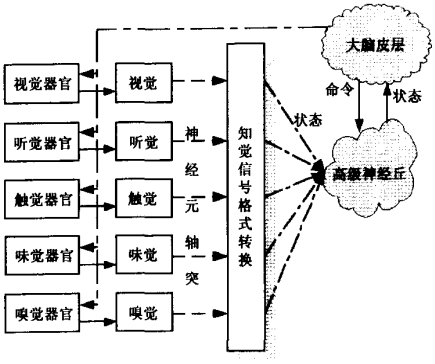

图 17 融合的生物认知模型

基于以上生物特性的融合模型,我们发现,信息融合并不是一个简单的单方向信息处理过程,而应该是一个具有反馈机制的闭环处理过程.基于此,在我们的第四代人机交互中提出了一种新的具有知识引导和反馈自调整特性的融合计算模型,如图 18 所示.

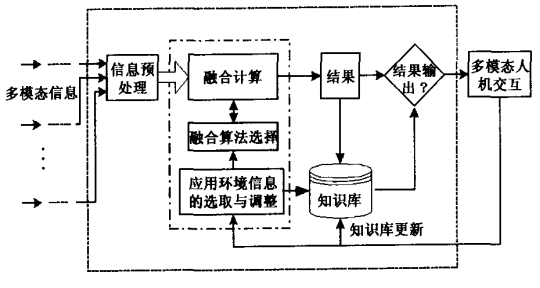

图 18 信息融合计算模型

这里主要模块的功能为:信息预处理模块主要对原始输入信息进行去噪等预处理,并将输入的多模态信息分配到合理的单元中.所有的融合计算都在融合计算模块中完成.融合过程针对于不同的应用环境,可以设计多个融合算法以供选择.同时可以在特征级和决策级上完成融合计算.融合算法的选择以及参数的设置由融合算法选择模块来控制.在实际应用中,一次往往只选取某一种算法进行融合计算.如果初次融合结果同知识库中的先验知识矛盾,则根据反馈选择其他的算法进行计算,该过程就是由算法选择模块来完成;同时,一个融合系统可以在不同的应用环境中运行,针对不同的上下文和对象信息,可以设置相应的环境信息,这个过程相当于将与该环境相关的先验信息输入到融合计算模块中参与计算.人类之所以能够对周围环境中的各种信号进行快速而准确的反应,正是由于他具备一个功能强大、结构合理的知识库.在多模态信息融合系统中,知识库同样具有不容忽略的地位.同时,知识库的内容要进行不断的积累和更新,该过程可以通过手工完成,也可以在运行过程中自动更新.知识库管理包括知识获取、恢复、存储、相关性分析、数据压缩和保护等,因而知识库管理是一个重要而又繁杂的任务.通过人机交互模块,用户可以向融合系统输入命令、请求以及表达对融合结果的满意程度等.利用这种方式,可以手动调整应用环境信息,从而实现对融合算法及融合参数的人工选择.此外,该模块是用户更新和管理知识库的手段,同时,这也是系统向用户展示融合结果的唯一途径.表 2 给出了考虑知识与反馈的身份识别实验结果,可以看到融合结果明显得到了提高.

表 2 考虑知识与反馈的身份识别实验结果

分类器 / 融合算法	本征脸 + 随机线	本征脸 + 语音	随机线 + 语音	本征脸 + 随机线 + 语音
Oracle 规则	96.5%	98.0%	99.5%	99.0%
单种融合算法下的最优结果	93.5%	92.0%	97.5%	94.5%
考虑知识与反馈的融合结果	94.5%	96.5%	98.0%	98.5%

4 实验平台

在国家自然科学基金重点项目"多功能感知机"研究基础上,对第四代人机自然交互系统的实现进行了研究,设计了第四代人机自然交互系统的实验平台和网络环境,并基于该平台实现了一个多功能感知信息服务系统,如图 19 所示.

图中智能代理由三维人脸表征,该人脸模型具有唇动合成功能,用户可以用语音与其交互.系统主要功能包括多通道信息的网络传输(立体图像传输),立体视觉系统,三维信息的获取和重建,AVR 三维重建及及漫游,以颐和园导游为任务域,建立的人机口语对话系统,基于对话的 Internet 信息查询,多媒体数据库和图像检索,人脸跟踪与识别,基于生物特性的视听信息融合,全面体现了第四代人机交互的四大特征.图 20 是视听信息融合的人脸识别界面.

图 19　基于 Agent 多功能感知信息服务系统

图 20　基于多分类器融合的人脸识别

5　结束语

本文给出新一代(第四代)人机交互的概念框架和特征,研究发展了其关键技术的解决方法,描述了实验系统平台的全面功能,可以为同类研究人员作参考.

致谢　感谢已毕业的博士、硕士,特别是陈芳、刘渭滨、丁凤、何伟、李红莲等.还要感谢中国科学院计算所高文教授的项目合作.

参考文献:

[1] Bolt R. Put that there: Voice and gesture at the graphic interface[J]. Computer Graphics, 1990, 14(3): 262 - 270.

[2] Fukumotoy M, Suenaga Y, Mase K. Finger-pointer: pointing interface by image porcessing[J]. Computer Graphics, 1994, 18(5): 633 - 642.

[3] Pittman J A, Smith I, Cohen P, Oviatt S, Yang T C. Quickset: a multimodal interface for military simulation[A]. Proc. 6th Conf. Computer-Generated Forces and Behavioral Representation[C]. Orlando, FL, 1996. 217 - 224.

[4] Cohen P R, Johnston M, McGee D, Oviatt S, Pittman J. Quickset: multimodal interaction for simulation set-up and control[A]. Proc. 5th Applied Natural Language Processing Meeting[C]. Washington, DC, 1997.

[5] Cheyer A, Julia L. Multimodal maps: an agent-based approach[A]. Proc. of the International Conf. on Cooperative Multimodal Communication[C]. Eindhoven. The Netherlands, Available at: http://www.ai.sri.com ~ oaa/ + "Bibliography", 1995.

[6] Vo M T, Waibel A. A multimodal human-computer interface: combination of gesture and speech recognition[A]. Proc. Inter CHI'93[C]. Amsterdam, Netherlands, Apr. 1993.

[7] Vo M T, Waibel A. Multimodal human-computer interaction[A]. Proc. of ISSD'93[C]. Waseda, Japan, 1993.

[8] Oviatt S L. Multimodal interfaces for dynamic interactive Maps[A]. In: Proc. of ACM Conf. on Human Factors in Computing Systems (CHI'96)[C]. ACM, 1996. 95 - 102.

[9] Oviatt S L. Multimodal interactive maps: designing for human performance[J]. Human-Computer Interaction, 1997, 12(1 - 2): 93 - 129.

[10] Adjoudani A, Benoit C. Audio-visual speech recognition compared across two architecture[A]. Proc. Eurospeech'95 Conf. [C]. Madrid, Spain, 1995. 2. 1563 - 1566.

[11] Akey M, Marsic I, Medl A, Bu G. A system for medical construction and education using multimodal human/machine communication[J]. IEEE Trans on Information Technology in Biomedicine, 1998, 2(4): 282 - 291.

[12] Pastoor S, Liu J, Renault S. An experimental multimedia system allowing 3-D visualization and eye-controlled interaction without user-worn devices[J]. IEEE Trans on Multiedia, 1999, 1(1): 41 - 51.

[13] Tsui K C, Azvine B, Djian D, Voudouris C, Xu L Q. Intelligent multimodal systems[J]. BT Technology, 1998, 16(3): 134 - 144.

[14] Tosa N. Neuro-Baby[DB/OL]. Available at: http://www.mic.atr.co.jp/tosa.

[15] Pentland A. Smart rooms[J]. Scientific American, 1996, 4: 54 - 62.

[16] Available at: http://www.is.cs.cmu.edu/js/[DB/OL].

[17] Available at: http://www.ri.cmu.edu/centers/vasc/[DB/OL].

[18] Aborzi H, Druin A. Designing storyrooms: Interactive storytelling spaces for children[R]. UMIACS-TR-2000-06, Available at: http://www.cs.umd.edu/hcil/, 2000.

[19] Thalmann N M, Kshirsagar S. Communicating with autonomous virtual humans[A]. Proc. of the 17th TWENTE Workshop on Language Technology[C]. Easchede, Universiteit Twente, Oct. 2000. 1 - 8.

[20] Kshirsagar S, Thalmann N M. Multimedia communication with virtual humans[A]. Proc. of Euromedia 2000[C]. Society for Computer Simulation International, Antwerp, Belgium, May 2000.

[21] Available at: http://www.miralab.unige.ch/[DB/OL].

[22] Gao W, Ma J, Chen X, Shan S, Zeng W, Yan J, Zhang H, Wu J, Wu F, Wang C. HandTalker: A multimodal dialog system using sign language and 3-D virtual human[A]. ICMI 2000[C]. Beijing, China, 2000. 564 - 581.

[23] 高文,陈熙林,马继勇,王兆其.基于多模态接口技术的聋人和正常人交流系统[J].计算机学报,2000,23(12):1253 - 1260.

[24] Brien P D, Nicol R C. FIPA - towards a standard for software agents [J]. BT Technol J, 1998, 16(3): 51 - 59.

[25] 王延江,袁保宗.软件 Agent 及其在多功能感知系统中的应用[J].信号处理,1999,15:441 - 444.

[26] Wang Yanjiang, Yuan Baozong. Face detection and location from color images using an evolutionary approach[J]. Chinese Journal of Electronics, 2000, 9(3): 270 - 273.

[27] Wang Yanjiang, Yuan Baozong. Face detection using evolutionary a-

gents[A]. The 5th Proceedings of IEEE International Conference on Signal Processing[C]. Beijing, 2000.2.1221 - 1224.

[28] Wang Yanjiang, Yuan Baozong. A novel approach for face detection from color images under complex background[J]. Pattern Recognition, 2001, 34(10): 1983 - 1992.

[29] Wang Yanjiang, Yuan Baozong. Novel segmentation method for face detection in complex background[J]. Electronics Letters, 2000, 36(3): 213 - 214.

[30] Chen Fang, Yuan Baozong. An approach to intellegent chinese dialogue system[A]. ICSLP2000[C]. Austrilia, 2000.595 - 598.

[31] He Wei, Ding Feng. Extracting entities for natural language dialogue system[A]. ICSP2000[C]. Beijing, 2000.2.764 - 767.

[32] He Wei, Yuan Baozong. Using context knowledge to improve the accuracy of recognition in voice browsing[A]. Proceedings of ICSP'02[C]. Beijing, 2002.

[33] He Wei, Yuan Baozong. Domain adaptation augmented by state dependence in spoken dialog dystems[A]. Eurospeech 2003[C]. Genva, Switzerland, 2003.1893 - 1896.

[34] Li Hongliang, Yuan Baozong. Research on voice browsing[A]. Proceedings of IEEE TENCON'02[C]. Beijing, 2002.1.473 - 476.

[35] 李红莲, 袁保宗. 一种文本相似度及其在语音识别中的应用[J]. 中文信息学报, 2003, 17(1): 60 - 64.

[36] Li Hongliang, Yuan Baozong. An improved support vector machines: NNSVM[J]. Chinese Journal of Electronics, 2004, 13(2): 321 - 324.

[37] Xing Qiang, Yuan Baozong. Unsupervised multiresolution image segmentation based on color moments [A]. Proceedings of ICSP'02[C]. 2002.2.584 - 587.

[38] Xing Qiang, Yuan Baozong. A new framework of CBIR based on KDD [A]. Proceedings of ICSP'02[C]. 2002.2.973 - 976.

[39] Hu Mingxing, Yuan Baozong, Tang Xiaofang. An new rubust fundamental matrics estimation based on genetic algorithm[J]. Computing and Informatics, 2001, 20(5): 471 - 486.

[40] Hu Mingxing, Yuan Baozong. Rubust estimation of trifocal tensor using messy genetic algorithm[A]. 16th IEEE ICPR 2002[C]. Quebec City, Canada, 2002.4.340 - 350.

[41] Liu W B, Yuan B Z. From actual reality to virtual reality[J]. Chinese Journal of Electronics, 2001, 10(2): 156 - 160.

[42] Liu W B, Yuan B Z. Superquadric based hierarchical reconstruction from 3D data[J]. Chinese Journal of Electronics, 2001, 10(1): 100 - 105.

[43] Rujie Liu, Baozong Yuan. A D-S based multi-channel information fusion method using classifiers' uncertainty measurement[A]. 5th Proceedings of IEEE International Conference on Signal Processing (ICSP' 2000)[C]. Beijing, 2000.1297 - 1300.

[44] 刘汝杰, 袁保宗, 唐晓芳. 用遗传算法实现模糊测度赋值的一种多分类器融合算法[J]. 电子学报, 2003, 30(1): 145 - 148.

[45] Rujie Liu, Baozong Yuan. Multiple classifiers combination by clustering and selection[J]. Information Fusion, 2001, 2/3: 163 - 168.

[46] Rujie Liu, Baozong Yuan. Information fusion modal and multi-perception system[A]. 5th Asia-Pacific Conference on Communications and 4th Optoelectronics and Communications Conference[C]. Beijing, 1999.985 - 989.

作者简介:

袁保宗 男, 1932 年生于江苏省吴江, 北方交通大学信息所教授, 名誉所长, 博士生导师, 目前, 他的主要研究方向包括语音信号处理、图像处理、计算机视觉、计算机图形学、虚拟现实以及人机交互技术等.