上下文场景中的多模态意图识别综述

——《设计认知及计算》课程作业

姓名： 张瑞升

学号： 220426

指导老师： 周小舟

日期： 2022.12.05

摘　要

近年来，随着大数据、人工智能等信息技术的飞速发展与多模态交互通道的不断融合，人机交互系统正在经历由单一模态向多模态、由显式命令执行向隐式意图推理、由基于当前状态的瞬时行为模式识别向基于上下文任务情境的连续交互意图感知等一系列转变历程。在该趋势导向下，计算机模拟人的认知感知机理、从上下文任务情境中解释并识别用户多模态的意图表达成为一项研究热点。

本文面向上下文场景中的多模态意图识别课题，从意图和上下文概念出发，阐释了本篇综述所涉及的概念范畴，提出了基于上下文场景的多模态意图识别框架，对当下基于上下文场景的多模态意图表征形式、特征提取方法、意图识别算法技术、意图识别应用领域进行了分析和归纳，并最终总结了该领域未来的潜在研究方向。

总体来讲，目前上下文场景下的多模态意图识别的研究涉及到两个全然不同的领域——对“人”的认知感知机理探究领域和对“机”的机器学习模式识别等人工智能领域，这需要建立对“人”和“机”的双向理解。更进一步说，人机交互研究需要从人的认知感知、计算机的响应决策双向着手，建立人的高级感知认知加工和计算机的高效智能系统的强耦合关系，构成双向感知、双向理解的人机融合共生体。

关键词：意图识别；多模态；上下文任务情境；人机交互

目 录

第一章 引言

1.1 课题背景

随着大数据、人工智能等信息技术的飞速发展与多模态交互通道的不断融合，人机交互系统逐渐由信息化向着智能化方向转变。在现实生活中，人与人之间的互动本质是多模态[1]的，我们使用语言、眼神、手势、肢体行为、生理信号等多种输出模态表达意图，并通过视觉、听觉、触觉、嗅觉等多个通道在上下文情境中理解意图。与之类似的，以自然交互、多模态融合为导向的智能人机交互系统要求计算机在复杂的交互情境中拥有类人的感知模式和认知处理机制，能够准确理解来自用户本源性认知与行为习惯引导下的多模态意图表达，并将意图解码成下一步行为指令，完成对用户意图的反馈，该过程如图 1-1 所示。正如心智理论[2]所描述那样，当人类能够快速感知推断出计算机的“思想状态”、同时计算机能够准确理解人类的意图线索时，人机系统的信任程度、交互效率、用户体验将大幅提升。在该目标导向下，计算机模拟人的认知感知机理、解释并识别用户多模态的意图表达甚至基于多模态意图线索推理预测出用户接下来的行为状态，是建立多模态融合的智能人机交互系统的关键环节之一。

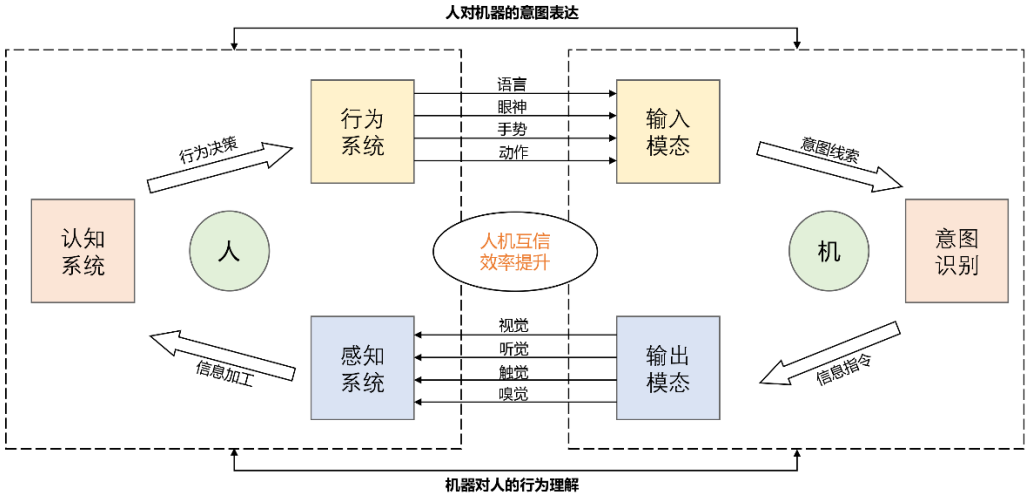

图 1-1 智能人机交互系统意图感知模式

近年来，智能人机交互系统正在经历由单一模态向多模态[3]、由显式命令执行向隐式意图推理[4]、由基于当前状态的瞬时行为模式识别向基于上下文任务情境的连续交互意图感知[5]等一系列转变历程，要从模糊的自然行为数据中更准确

地推理人类意图，不仅需要更多模态的生理行为意图线索，还需要上下文任务情境的参与。上下文在不同领域、不同学科下有着不同概念与范畴[6]，在智能人机交互系统中，上下文囊括了用户所执行的任务事件与本体状态、交互系统的物理环境与语义情境、计算机的决策推理状态等多种信息，它随着任务情境和交互过程不断动态发展，影响着意图推理过程中实体与实体、实体与语义之间的关系变迁，决定着机器对用户、环境和任务的感知与解释[7]。上下文与意图线索之间密切联系，同一意图线索在不同上下文任务情境下往往传递着不同含义[8]，因此，多模态意图识别推理离不开上下文任务情境的指导，了解上下文任务情境是决定计算机是否正确解释用户意图的关键所在。目前，基于上下文任务情境的多模态意图识别已经在日常任务[9]、战场作战[10]、自动驾驶[11]、人机协作[12]、人机交互[13]等多个领域开展着广泛探索。

1.2 组织框架

本篇综述共包含六章，各章节组织框架如图 1-2 所示。

第一章，引言。本章阐释了上下文场景中的多模态意图识别课题背景，说明本篇综述的文献搜集整理方法，并梳理全篇组织框架。

第二章，上下文场景中的多模态意图识别概念与架构。本章阐释了课题相关的意图、上下文等基本概念，指出本文所涉及的意图与上下文的特定范畴，最终整理了基于上下文场景的多模态意图识别整体框架。

第三章，上下文场景中的多模态意图表征形式与特征提取方法。本章阐释了人机交互系统中交互意图的多模态表征形式、各模态的特点及各模态意图特征提取方法。

第四章，上下文场景中的多模态意图识别算法技术。本章介绍了已知意图输入特征前提下适用于多模态意图识别的相关算法技术，并总结了各算法技术的优缺点与适用范围。

第五章，上下文场景中的多模态意图识别应用领域。本章归纳了基于上下文场景的多模态意图识别技术在各领域的具体应用情况。

第六章，总结与展望。本章阐释了上下文场景中的多模态意图识别研究对人机系统的意义，并总结了该领域未来的潜在研究方向。

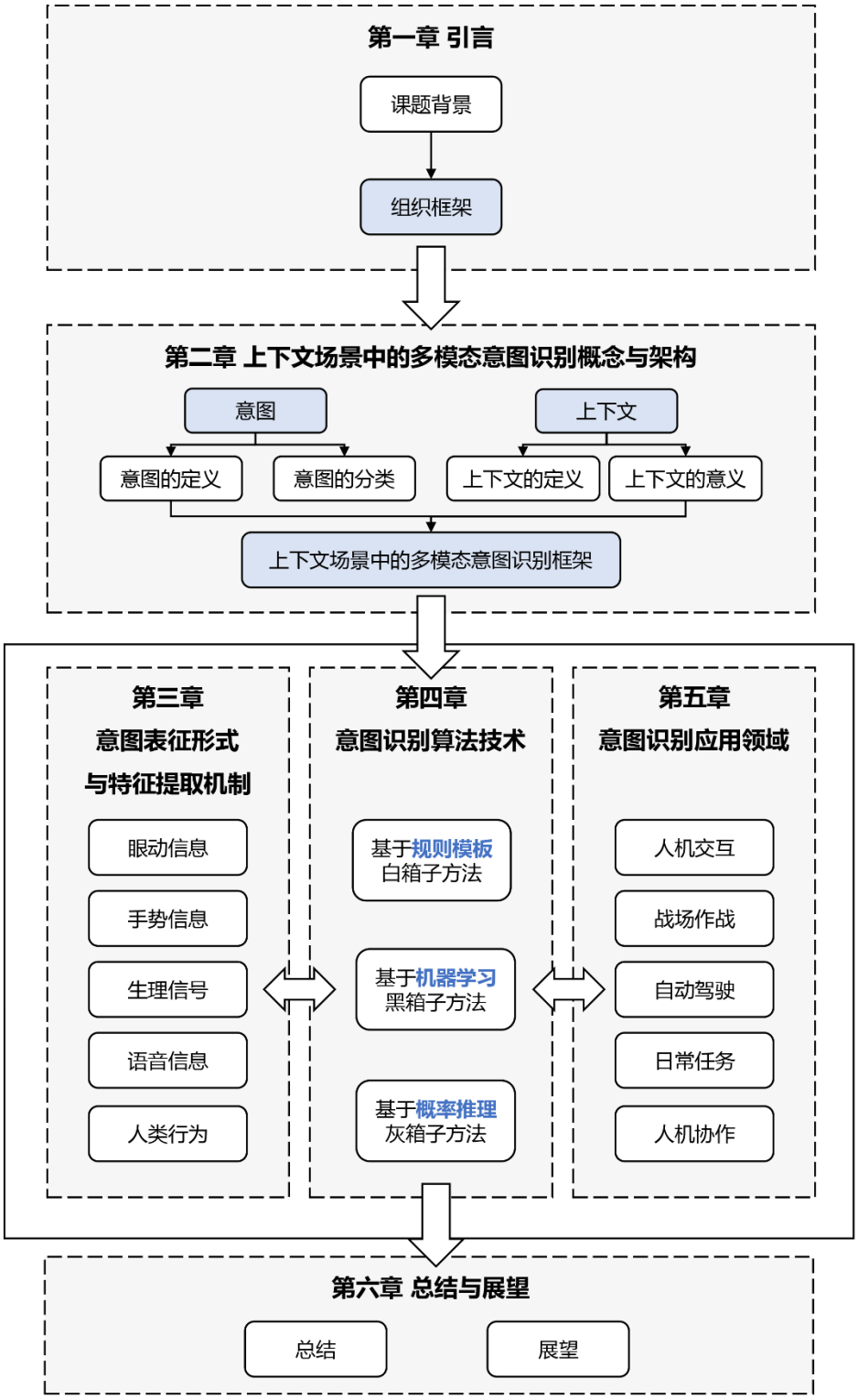
第一章 引言
课题背景
组织框架
第二章 上下文场景中的多模态意图识别概念与架构
意图
上下文
意图的定义
意图的分类
上下文的定义
上下文的意义
上下文场景中的多模态意图识别框架
第三章
意图表征形式
与特征提取机制
眼动信息
手势信息
生理信号
语音信息
人类行为
第四章
意图识别算法技术
基于规则模板
白箱子方法
基于机器学习
黑箱子方法
基于概率推理
灰箱子方法
第五章
意图识别应用领域
人机交互
战场作战
自动驾驶
日常任务
人机协作
第六章 总结与展望
总结
展望

图 1-2 本文组织框架

第二章 上下文场景中的多模态意图识别概念与架构

广义上的意图识别是指人或智能体根据用户提供的明确或隐含的信息中识别用户的目标与期望的过程，而本文中的意图识别则聚焦于人机系统，强调计算机对人的意图感知，并且与上下文场景息息相关。本章将阐释意图与上下文场景的基本概念及本文中的意图与上下文场景的特指范畴，并最终提出基于上下文场景的多模态意图识别研究的整体架构。

2.1 意图

2.1.1 意图的定义

意图是一个人对于执行特定行动的承诺，是由目标期望到具体行动的纽带[14]，它在行动之前形成，受个人实现特定目标的愿望驱动[15]，并流露出用户自然表现出的行为线索。心智理论中，科学家们已经通过黑猩猩实验证明人类天生具备理解他人思想的能力，能够从行为线索中预测、理解他人明确或隐含表达的意图[2]。Blakemore 等[16]基于心理物理学和神经生理学研究基础，模拟生物运动对大脑的作用，并揭示人类模拟自己观察到的他人行为并将其映射到自身的意图表达上的意图行为产生机理。在心理学理论里，意图起源于人的期望，但又与人的期望存在差异，二者主要区别在于将意图或期望转换成现实行为的具体程度[17]。

在人机交互系统中，意图是用户交互意愿的表达，是建立人的模糊认知空间到计算机的精确计算空间的映射手段[18]。交互意图是明确的或隐含的，其中：明确的交互意图将通过交互界面直接输入到人机交互系统中；隐含的意图则涉及到用户内部认知感知加工机制，需要根据用户的自然表情、手势、语言、眼球移动等意图线索提示来进行推断。明确的交互意图保证了人机系统之间沟通的准确性，而隐含的交互意图注重于实现人机系统交互的自然性，其最终目的是使计算机能够像人类一样理解和预测用户的行为与目的。无论是在现实世界的互动状态还是虚拟世界的交互行为，人无时无刻不在表达着自己的意图线索[9]，如何根据用户自然行为下的模糊行为线索数据推理出用户的当前交互意图，是智能人机交互和自然人机交互领域亟待解决的难题。

2.1.2 意图的分类

人机交互系统中的意图复杂多样，给予交互意图确定的分类形式将有助于完成对不同类型交互意图的界限规划，以便后续根据不同意图特性进行上下文用户意图建模。Lukander 等[19]认为人机交互系统中的意图隐含着人的注意力分布状况，其根据注意力建模方式将交互意图分为自下而上的意图与自上而下的意图两类：自下而上的意图受交互场景刺激引导，根据交互系统中的刺激结构可以预测出用户意图的空间分布；而自上而下的意图受用户当前执行任务流程引导，根据用户意图线索可推理出当前用户的任务状态。Schmidt[6]将人的交互意图分为显式意图和隐式意图：显式意图依赖于用户明确的意图表达，体现用户对于计算机系统的直接命令，该类意图直观明确、易于理解，但受制于局限的输入模态且占用用户较高的认知工作量[20]；隐式意图注重计算机主动感知用户自然行为线索并推理当前意图，体现计算机系统对于用户的自动响应，该类意图不需要用户明确的命令传达和先验知识，但复杂性高、难以准确识别。马丽莎等[21]基于眼动注视模式将隐式意图进一步分为非目的性意图和目的性意图：非目的性意图表达用户在非任务状态下对于感兴趣目标的扫描行为；目的性意图表达用户在特定任务下的特定目标搜索行为。O'Connell 等[22]根据自下而上的刺激方式，将自下而上的意图细分为基于刺激显著性的外源性意图和基于刺激内容驱动的内源性意图，而自上而下的意图为基于任务目的驱动下的内源性意图。Karaman 等[4]根据上下文任务情境将交互意图分为现实任务意图与虚拟任务意图：现实任务意图聚焦于现实世界自然交互行为的分析推理；虚拟任务意图聚焦于用户与计算机系统交互行为的意图识别。此外，依据意图分类模型与意图本体的连续离散特性，意图推理可分为分类和回归两类问题：意图分类根据用户输入行为线索，从多个可能意图状态中检测相应意图；意图回归则赋予意图线索具体数值或指标，以提升意图识别准确程度和响应精度。依据用户当前交互任务特性[23-25]，意图还可以分为系统控制意图、符号输入意图、选择指向意图、选择确认意图、修正意图、操纵意图、创建意图等。本节所涉及的意图分类整理如表 2-1 所示。

表 2-1 人机交互系统中的交互意图分类

分类依据	意图类型	意图定义
注意力建模理论[19]	自上而下的意图	基于用户执行交互任务引导的意图
	自下而上的意图	基于交互场景刺激引导的注意力意图
人机交互显隐性[6]	显式意图	明确意图，由用户对计算机的直接命令
	隐式意图	隐含意图，需计算机自动识别用户期望
基于眼动信息的隐式意图特征[21]	目的性意图	非任务状态下的自然注视意图
	非目的性意图	特定任务状态下的视觉搜索意图
意图驱动形式[22]	外源性意图	基于刺激显著性引导的外部驱动意图
	内源性意图	基于刺激内容与自身任务引导的内部驱动意图
上下文任务情境[4]	现实任务意图	现实世界自然交互行为所产生的意图
	虚拟任务意图	用户与计算机系统交互行为所产生的意图
意图分类模型与本体连续离散特性	分类意图	离散意图，体现为各类意图状态的推理分析
	回归意图	连续意图，聚焦于意图识别的程度与精度
交互任务特性[23-25]	系统控制意图	用于系统控制任务的交互意图
	符号输入意图	用于文本语音等符号输入的交互意图
	选择指向意图	用于目标选择任务的指向意图
	选择确认意图	用于目标选择任务的确认意图
	修正意图	用于三维物体修改的交互意图
	操纵意图	用于三维物体操纵的交互意图
	创建意图	用于三维物体创建的交互意图

2.2 上下文场景

2.2.1 上下文的定义

上下文的概念在不同领域有着不同的说法，本文特指人机交互系统中的上下文场景。在人机交互系统，上下文场景是用户在交互过程中所处的整个任务情境，囊括人机交互系统中“人-机-环”各个组成部分。人机交互系统中的上下文场景根据信息来源可分为：用户上下文、计算上下文与环境上下文三个部分[57]。如图

2-1 所示，用户上下文场景涉及用户在交互过程中的执行任务即总体期望、用户生理心理的本体特性、用户在交互系统中的时空关系、用户的社会背景等信息；计算上下文涉及交互系统中计算机所捕获到的多模态数据输入情况、本机信息处理能力、交互环境网络连接状况、计算机历史决策推理状态等信息；环境上下文涉及人机交互系统总体情境，依据特性分为物理环境、功能情境、语义情境、情感情境等。Wei 等[58]还根据上下文场景的显隐特性，将其分为显式上下文与隐式上下文：显式上下文涉及用户在特定任务下的工作状态、上下文语义信息、人类行为生理特征信息、人类行为上下文运动学状态、多模态意图的外显特征等；隐式上下文则涉及交互场景的功能情境、情感背景和社会文化背景等。

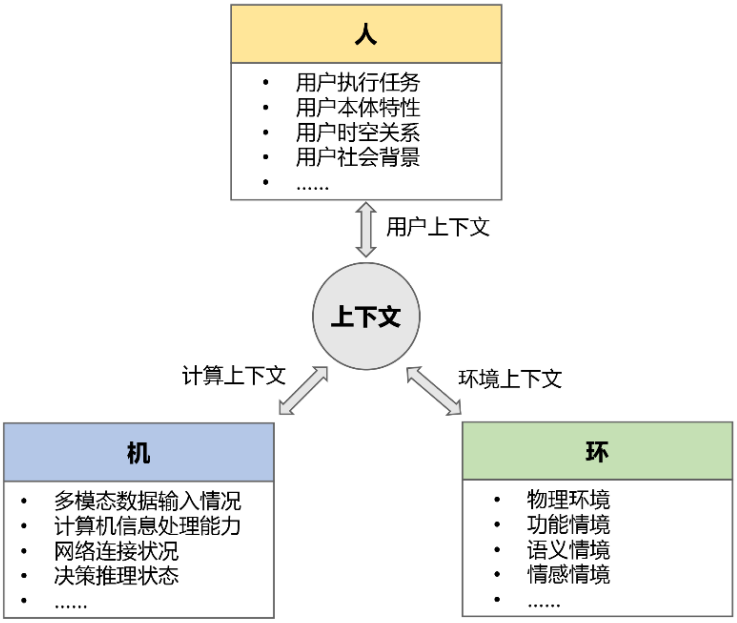

图 2-1 人机交互系统中的上下文场景范畴

2.2.2 上下文对于多模态意图识别的意义

上下文场景具有动态性、依赖性、时序性的特点[57]，即：基于上下文场景的建模过程将不断动态更新大量不同语义级别的上下文信息源；其所捕获的上下文信息存在着实体、属性、语义间各种复杂依赖关系，这些信息之间并非独立无关而会互相影响；上下文场景能够通过了解“过去”所存在的先验信息，以预测“未来”的状态分布，时序性是决定其推理能力的重要特征。基于上述特点，上下文场景对于人机交互系统的意图识别问题的意义为：

1.动态性：人的多模态意图表征常为动态过程，从静态维度感知容易造成意图识别过程中的“断章取义”问题。

2.依赖性：用户的各模态意图表征与上下文场景具有强烈的依赖关系，不同任务情境下意图表征语义大不相同。

3. 时序性：用户意图是过去行为、过去决策、过去意图的延续，以动态上下文场景中的先验知识为桥梁，有利于计算机逐步理解各层次的意图语义，跨越离散情形下的语义鸿沟，从而逼近对于用户意图的正确理解。

2.3 基于上下文场景的多模态意图识别框架

基于上下文场景的多模态意图识别框架如图 2-2 所示。上下文场景中的多模态意图识别经历如下几个阶段：一、数据提取阶段：记录各个模态下的意图表征原始数据，从中提取出有效的意图特征；二、意图识别阶段：基于上下文的任务情境、人的“感知-认知-决策-行为”模型、输入意图特征的模态特点等，选择合适的意图识别分类技术；三、技术应用阶段：将建立的意图识别模型投入到具体的应用场景，检验意图识别效果。接下来三章将对各个阶段进行详细阐释。

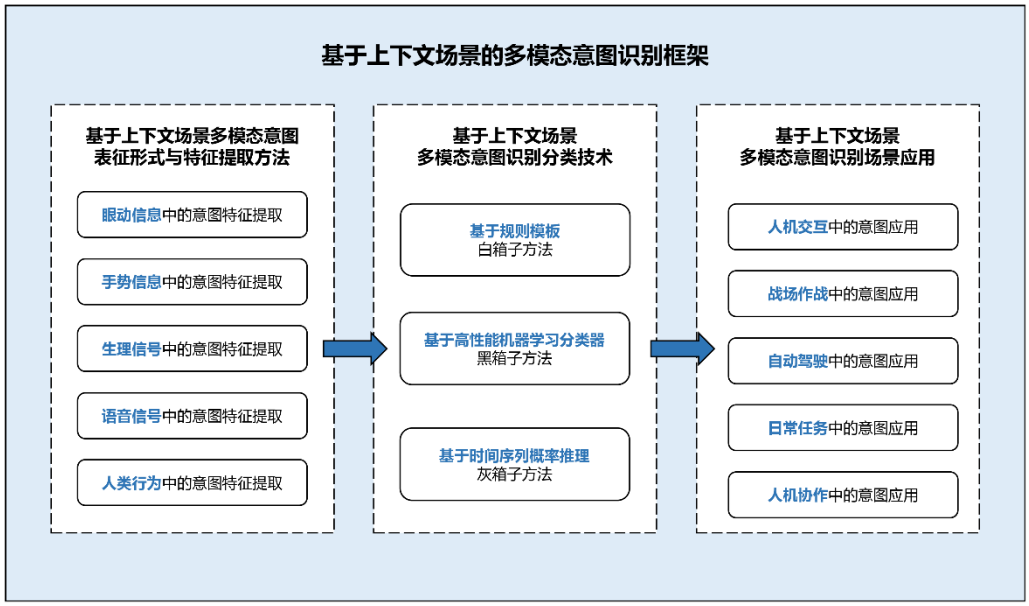

图 2-2 基于上下文场景的多模态意图识别框架

第三章 上下文场景中的多模态意图表征形式与特征提取方法

人机系统中的交互意图以多模态的形式表征，用户自然地通过眼神、言语、手势、肢体动作表达自己的交互期望，并且这一过程常伴随生理信号、注意力等多种形式的生理心理特征变化。随着传感技术的发展，交互意图线索在各类模态下的外显特征得到测量与量化，研究者们开始尝试通过这些意图线索的外显特征来推理分析出人的交互意图。受制于传感设备测量方式、精确程度、稳定性等方面的限制，各类模态下的意图表征各有其优势与不足，融合多种模态表征形式的交互意图推理成为建立人机系统自然有效沟通的必要手段。本章将对眼球运动[28]、生理信号[29]、手势信息[30]、语音信息[31]、人类行为[32]等常用意图表征模态及各模态特征的特征提取方法进行简要阐释。

3.1 基于眼球运动的意图

3.1.1 基于眼球移动的意图表征形式

眼睛是人视觉感知的重要通道。人在执行交互任务时常通过感知系统的各感官通道在周围环境中感知信息，并根据这些信息来指导自己的行动，其中，视觉通道最为重要，人感知到的外界信息 80%以上从视觉通道获取[33]。在视觉感知过程中，由于视网膜细胞的不均匀分布，个体必须要将感兴趣目标落在视网膜中央凹区域上以最大限度提取细节，因此人的眼睛必须通过不断移动来观察周围环境。

在认知心理学中，眼睛移动被视为视觉注意力的表征[34]，其受到认知、期望、任务、记忆等自上而下认知因素影响并反映当前环境中自下而上的感官刺激。在过去研究中，研究者们常将眼睛作为感受器来探究眼睛与注意力、人的感知相关研究，而近期越来越多的研究表明，人的眼睛在为行动计划收集视觉信息的同时，也可作为效应器来提供信息，以推断用户的意图和下一步的计划。眼心假说[35]认为直接的眼球运动模式反映复杂的潜在认知过程，个体眼睛所观察的位置与思考具有强的关联性。Castiello [36]通过心理学实验证明人可以通过观察他人的目光来判断对方的交互意图。Yarbus[37]早在上世纪七十年代开始探究人在感知复杂物体时的眼球移动，其结果表明在不同的任务情境下观察同一物体时人眼的视线轨迹

存在着巨大的差异，而根据眼动信息推理任务情境并提取用户意图的过程被称为逆 Yarbus 过程[19]。

在人机交互系统中，眼球移动最开始以显式输入方式应用于命令界面，然而这种强迫用户有意识地采用不自然的注视行为进行系统交互的方式会造成较高的认知负荷且存在 Midas 接触[38]、交互粒度过粗等问题，故常应用于渐冻症等特殊人群使用。随着眼睛与注意力意图间的密切联系得到广泛论证，眼球运动逐渐以隐式输入方式运用在非命令的智能人机界面中[39]，计算机通过持续观察用户自然行为下眼球移动的各项特征从中推断出意图，眼球运动已成为意图表征最主要的模态之一。

3.1.2 眼球移动相关意图特征提取方法

眼球移动中的意图特征提取方法基于现代眼动追踪系统，一般采用基于红外相机的计算机视觉技术来完成眼球运动追踪和数据收集，其基本原理为：向人眼发射红外线，通过分析眼睛角膜表面光源反射情况确定人的视线方向，同时通过图像处理识别人眼瞳孔、虹膜、巩膜等视觉特征变化。Kar 等[40]总结了眼动追踪系统中所采用的具体算法。基于眼动仪的使用场景，现代眼动追踪系统可分为桌面式眼动追踪系统与可穿戴式眼动追踪系统两类：桌面式眼动追踪系统将眼动仪放置于距离用户一定距离的地方（例如桌面）观测眼球运动，常应用于驾驶模拟等固定实验场景；可穿戴式眼动追踪系统将眼动仪集成在~~眼睛~~眼镜或头盔上，常应用于虚拟现实等移动式实验场景。

眼球运动依照运动特点可以分为注视、扫视（也称眼跳）、平滑追踪、~~注视、~~眨眼等运动状态，这些运动中的各项信息记录着观察者及观察目标相关的异常丰富的信息来源，是所有基于眼动信息的意图推理和分类问题的研究基础[41]。眼球在各运动状态下表征意图的眼动特征总结如表 3-1 所示。扫视运动表示眼睛从一个注视点向另一个注视点的极短时间内快速跳跃，其常用度量特征涉及持续时间、幅度、潜伏期与速度方向等，常用于监测用户意图是否出现转移[42]。注视运动表示眼睛将中央凹视野区域在目标上保持一定时间以获取更多细节，其常用度量特征涉及持续时间、注视位置等，被广泛应用于推理意图目标所在方位[43]。眨眼运动的常用度量特征涉及眨眼持续时间、眨眼频率、眼睑闭合度等，用于检测用户

疲劳状态。此外，Coutrot 等[41]将眼睛的感兴趣区域（Area Of Interest, AOI）用于意图识别场景，通过记录由注视点聚类生成的 AOI 范围、数量及注视顺序以分析用户对视觉元素的扫描路径，并基于扫描路径推理用户当前的任务意图。Jang 等[44]建立了不同光环境下的瞳孔反应基线模型，将瞳孔特征用于用户视觉搜索意图识别任务。然而，人眼瞳孔特征极易受到视觉刺激大小及光环境的影响，常需要结合眼球运动的其他特征用于意图识别系统。

表 3-1 眼动信息中的意图特征

眼球运动状态	眼动特征具体参数
注视运动	注视持续时间、注视位置[43]
扫视运动	扫视持续时间、扫视幅度、扫视潜伏期、扫视速度与方向[42]
眨眼运动	眨眼持续时间，眨眼频率，眼睑闭合度
其他它	瞳孔扩张收缩程度[44]
	AOI 停留时间、AOI 范围、AOI 数量、AOI 注视顺序[41]

3.2 基于手势信息的意图

3.2.1 基于手势信息的意图表征

手是人执行运动和操作的主要通道。三维人机交互系统中，自然手势是用户向计算机阐明自身显式意图的主要手段[45]。基于手势信息的意图表征以最符合人本源性认知的非接触式交互手段完成人机对话，它保证了交互速度、交互舒适度与交互沉浸感，被广泛应用于三维用户界面。然而，人的手势具有较大的模糊性，不同社会背景的人对同一手势存在语义认知差异，即使同一个人也很难做出完全相同的手势，这为手势动作与交互意图间的匹配带来挑战[46]。针对上述问题，基于手势信息的意图识别通常采用专家系统[41]与启发式手势[48]两种策略：专家系统要求人机系统中用户扮演专家的角色，在表达和理解手势意图、手势语义等方面得到充分训练；启发式手势则注重于计算机基于已知手势意图去推理与理解用户的未知手势意图。此外，基于手势信息的意图表征常出现在与意图形成的最后阶段，一般伴随决策和行为同时产生[49]，因此，基于手势信息的意图表征一般仅应用于意图识别，对于人机系统中的意图推理和意图预测等环节，手势信息必须协

同意图形成的早期表征模态（例如眼球运动、生理信号等）共同完成。

3.2.2 手势信息相关意图特征提取方法

手势信息中的意图特征提取方法聚焦于手势识别与手势关键点信息提取，主要基于非接触式计算机视觉技术和接触式的可穿戴传感器实现。基于手势信息相关意图特征如表 3-2 所示。基于计算机视觉的手势意图特征提取方法采用普通 RGB 相机或深度相机（如 Leap Motion，Kinect）识别人手部骨骼关键点[50]，并根据关键点数据重建手部二维三维位姿状态，进而完成对手部信息的采集。此类意图特征提取方法成本低廉且直观自然，但受制于交互环境、手部遮挡等问题。基于可穿戴式传感器的手势意图特征提取方法采用 IMU 等惯性传感器测量人手部肌肉运动、手部速度、手部加速度等行为数据完成对手部意图相关特征的采集。此类意图特征提取方法更为准确，但侵入性较强，不利于人机系统的自然互动。

表 3-2 手势信息中的意图特征

手势识别方式	手势特征具体参数
计算机视觉技术	手部骨骼关节点世界坐标，关节点间相对坐标，手部位姿[50]
	手部速度、手部加速度等运动状态
可穿戴传感技术	手部速度、手部加速度等运动状态
	手部肌肉电信号[51]

3.3 基于生理信号的意图

3.3.1 基于生理信号的意图表征

生理信号是表征人体生理变化最快速的方式，是不同于眼睛、手势等外显行为的内隐式意图表达。人的意图起源于大脑的认知和决策活动，且往往伴随着个体为了达到期望所采取的行为活动，这些活动都会造成人体生理电信号变化。通过收集并解析来自不同组织、器官的神经系统生理电信号，可以解释个体当前的认知感知状况，进而理解人的意图。常用的生理信号意图表征形式有脑电[52]、肌电[53]、眼电[54]等。Novak 等[49]探索了脑电、眼电、眼动追踪、肌电、用户行为等不同技术对于预测人体到达运动目标的性能，其总结了不同生理信号预测行为意

图时在时空关系上的差异。虽然已经有很多基于单一模态生理信号的意图识别研究，但各模态的生理信号仍然存在一定的局限性。例如，脑电信号存在微弱性、信噪比低的特点，且容易受到用户动作与精神状态干扰[55]；肌电信号随着肌肉疲劳程度的提高而显著性降低，进而导致意图识别准确率降低[56]。此外，各模态生理信号并非独立存在的，Xi 等[57]证明脑电和肌电信号之间存在显著关联性。合理利用各模态生理信号的关联性，融合多模态的生理信号意图表征有利于探究人的意图形成机理和帮助机器更准确理解用户意图。

3.3.2 生理信号相关意图特征提取方法

生理信号相关意图特征包含脑电相关意图特征、肌电相关意图特征与眼电相关意图特征。脑电相关意图特征一方面采用脑电帽等头皮电极装置、基于叠加平均原理捕获事件相关电位(Event-related Potential, ERP)，实现对意图特征的时间维度的解析；另一方面采用近红外光谱仪或功能磁共振成像仪等装置捕获脑部活跃区域分布，实现对意图特征的空间维度解析[58]。肌电相关意图特征采用各类肌电传感器记录肌电信号（Electromyography, EMG），并通过模糊神经系统等方法完成对复杂 EMG 信号的分类[59]。眼电相关意图特征记录研究周围皮肤表面的电极变化，基于眼电信号（Electrooculography, EOG）表征意图特性[19]。

表 3-3 生理信号中的意图特征

生理信号表征方式	生理信号意图特征的具体参数
脑电	ERP 信号，脑部活跃区域分布
肌电	EMG 信号
眼电	EOG 信号

3.4 基于语音信息的意图

3.4.1 基于语音信息的意图表征

语言是人与人沟通最常用的表征方式，是用户意图表达最为直观明确的人机交互方法。随着自然语言处理(Natural Language Processing, NLP)领域的蓬勃发展，研究者们开始尝试摆脱对传统图形用户界面和物理按钮控制方式的依赖[60]，

寻求更普适更自然的意图表达形式，语音意图表征便是解决方案之一。基于语音的意图表征鼓励用户使用更简单、更自然的语言向计算机传达自身意图，计算机通过自然语言识别系统和自然语言解释系统完成用户意图的语义理解，并建立意图语义与交互实体之间的联系[61]，实现用户当前意图识别和语音交互响应。然而，基于语音信息的意图表征同样具有较大的模糊性和不确定性，不同语种、不同语境下的语音意图表达将存在显著差异，且容易受到任务场景与物理环境干扰[60]。因此，基于语音信息的意图表征常用于较为局限、稳定的人机交互系统中。

3.4.2 语音信息相关意图特征提取方法

语音信息相关意图特征主要涉及用户音频信号，直接使用麦克风设备记录用户语音信息。

3.5 基于人类行为的意图

3.5.1 基于人类行为的意图表征

人类行为是宏观角度上个体甚至群体的意图表征形式。基于人类行为的意图表征聚焦于个体的面部表情[62]、肢体动作[63]和群体的社会活动[64]，已经被广泛用于人和广义智能体的交互系统中。在人机交互领域，人类行为一方面作为交互意图输入通道提供人机交互的意图线索[65]，另一方面作为个体情绪、群体情感的意图表达以辅助完成人机情感化互动[66]。作为交互意图线索，人类行为适用于较远距离的人机交互情形，但仍然受制于模糊的意图表达、有限的交互空间、人体生物力学负荷限制，应用场景较为局限；作为情感意图线索，人类行为适用于广义智能体对用户情绪意图理解与社会活动意图解释，但需要复杂的意图特征输入与长时间的识别响应，且强烈依赖于对于目标个体的熟悉程度、共享社会文化背景等先验知识[67]，在意图识别精度与速度上表现不佳。

3.5.2 人类行为相关意图特征提取方法

人机交互系统中的人类行为相关意图特征主要涉及个体肢体动作、面部表情。其中，个体肢体动作的特征提取与手势意图特征提取方法类似，均基于计算机视

觉或是惯性传感器识别个体的肢体关节点位置，从而完成对个体位姿的表征[63]。面部表情的特征提取常基于计算机视觉技术或热成像技术等手段完成对人脸五官特征扫描，提取不同情绪下的面部动力学特征[68]，进而实现个体情绪识别。基于人类行为的相关意图特征总结如表 5-4 所示。

表 3-4 人类行为中的意图特征

人类行为类别	人类行为特征具体参数
肢体动作	肢体关节点位置、人体位姿[63]
面部表情	面部动力学特征[68]

第四章 上下文场景中的多模态意图识别算法技术

基于准确有效的多模态意图特征输入，计算机需要恰当算法完成对这些连续、非确定性的输入意图特征分析，并实现用户交互意图的识别与解释。本章将主要综述上下文场景中常用的多模态意图识别算法技术，依据这些技术所采用意图识别模型的可解释性，将其分为基于规则模型的意图识别技术、基于高性能机器学习分类器的意图识别分类技术、基于时间序列的概率推理意图识别技术。

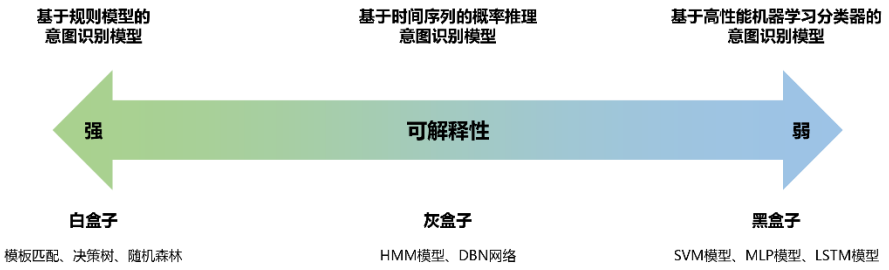

图 4-1 上下文场景中的意图识别模型基于可解释性分类

4.1 基于规则模型的多模态意图识别技术

基于规则模型的意图识别技术源于人对客观规律的数学建模，根据少量意图特征数据输入便确定模型参数。此类模型算法工作机制是透明的，所有规则由人工设计或训练得到，具备较强的解释性。然而，由于模型较为简单，基于规则模型的意图识别技术的预测能力通常较为局限，无法对输入特征的内在复杂依赖关系进行建模，一般用于简单的意图识别分类问题[46]。基于规则模型的多模态意图识别常用模型包含：模板匹配法[69]、决策树算法[70]、随机森林[71-72]等。

4.1.1 模板匹配法

模板匹配法是最原始、最简单的模式识别方法，其针对不同类别的分类结果制作模板库，通过对比当前意图特征输入与模板库中的相似程度，完成意图识别。其简化模型如图 4-2 所示。Elakkiya 等[69]将模板匹配法用于手势意图识别，通过将手势与手语标识进行匹配完成计算机对用户手语表达的实时理解。然而，当模

板过多或模板之间相似性过强，该方法的意图识别准确率将大大降低，故此方法仅适用于明确简单的显式意图识别。

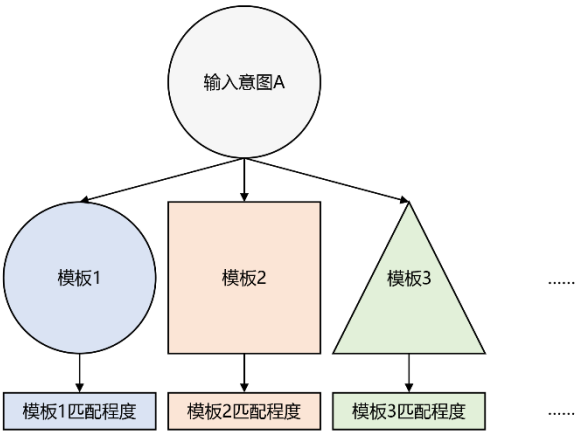

图 4-2 模板匹配法意图识别简化模型

4.1.2 决策树算法

决策树算法是一种树形结构的意图分类方法，其最简单模型如图 4-3 所示，每个树节点都表示对一属性或输入特征的判断，判断规则通过模型训练得到，每个分支代表判断结果，最后每一叶节点代表最终分类结果。Zhou 等[70]人将决策树算法运用在战场空中目标意图预测中，其首先采用 LSTM 网络收集目标状态数据并预测未来状态信息，之后利用决策树算法从不确定、不完备的先验知识信息库中提取状态规则来获得空中目标的战场意图。然而，决策树算法在面对庞大数据量时易出现过拟合问题，且忽略了输入特征之间的相互关联性，故常用于差别显著的状态意图识别。

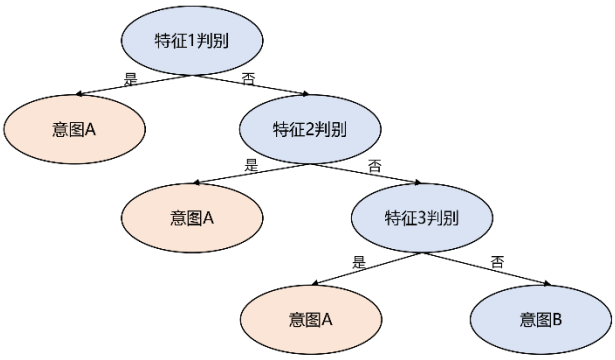

图 4-3 决策树算法意图识别简化模型

4.1.3 随机森林

随机森林在决策树算法的基础上优化决策方式，相较于决策树创建规则进行分类的形式，随机森林采用随机选择特征构建决策树，通过评估各决策树输出结果的一致性完成意图分类，其简化模型如图 4-4 所示。由于大量随机产生的不相关决策树协同工作，随机森林可有效避免传统决策树算法过拟合问题。Boisvert 等[71]将随机森林算法引入任务意图预测问题中，基于眼动注视特征输入预测用户未来的任务状态。Liao 等[72]将随机森林算法应用于现实导航任务意图预测中。然而，由于随机属性的引入，随机森林并不属于白盒子范畴，其响应函数高度非线性，模型可解释性较低，且作为描述型模型也同样仅适用于意图识别分类问题。

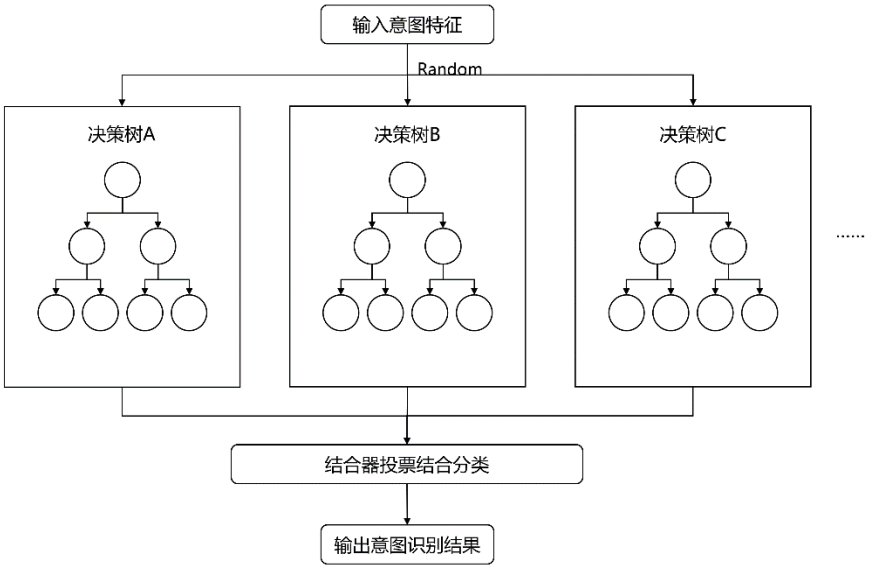

图 4-4 随机森林算法意图识别简化模型

4.2 基于高性能机器学习分类器的多模态意图识别技术

基于高性能分类器的意图识别技术采用隐变量机器学习方法，通过不可见的随机过程与神经网络来对意图特征输入进行建模。此类模型由人工设计规则转变为从数据中学习规则，可以执行复杂意图任务特征建模。然而，由于这类模型内部工作机制完全空白，我们无法估计每个输入特征对模型预测结果的重要性，也难以理解各输入特征之间的相互作用，可解释性较差。此外，机器学习模型高度依赖数据集，良好的输入特征是此类技术效果优劣的重要因素。常用的高性能机器学习分类器有支持向量机（Support Vector Machine, SVM）模型[73]、多层感知

机（Multilayer Perceptron, MLP）[27]、长短时神经网络(Long Short-Term Memory, LSTM)[74-75]。

4.2.1 支持向量机模型

支持向量机是一类按监督学习方式对输入特征进行二元分类的广义线性分类器，其特征为计算速度快、准确率较高,是针对分类问题采用的最广泛的机器学习方法。Huang 等[9]将 SVM 模型应用于日常任务下的任务意图预测问题，通过输入眼动注视等意图特征实现制作三明治过程各任务环节的预测。Koochaki 等[73]同样将 SVM 模型应用于日常任务意图预测，其首先采用聚类算法和 CNN 神经网络从眼动注视等意图特征提取出 AOI 区域，再通过 SVM 模型根据预处理后的 AOI 特征完成意图分类任务。然而，SVM 模型不适用于大样本、过高特征维度的分类场景，在使用之前常需要对输入特征进行预处理。

4.2.2 多层感知机

多层感知机是经典的前馈型全连接神经网络，由输入层、输出层以及至少一个隐藏层组成，其特征为学习能力强，适用于普遍意义上的线性、非线性意图识别分类问题。Wei 等[27]将 MLP 模型运用于机器人的意图识别领域，经 LSTM 模型处理后的人类行为意图特征及环境特征被输入多层感知机，实现对动态人类行为的意图理解。然而，多层感知机在实现之前需要网络结构、权值等大量参数，导致学习时间过长，且易陷入局部最优问题。

4.2.3 长短时神经网络

长短时神经网络是一种时间循环神经网络，其特点是引入时序信息，能够基于时间先验信息辅助推理过程，被广泛应用于处理较长时间序列下的意图识别问题。Wang 等[74]将 LSTM 模型应用于长时间的驾驶环境，通过记录踏板与陀螺仪的运动特征识别驾驶员制动意图。Teng 等[75]使用类 LSTM 模型门控循环单元进行战场内空中目标战斗意图识别方法，引入双向传播机制和注意机制，建立空中目标特征与目标作战意图的映射关系。然而，LSTM 模型需随着时间推移进行顺序处理，复杂的单元网络结构使得 LSTM 训练过程中产生较大的运算量。

4.3 基于时间序列概率推理的多模态意图识别技术

基于时间序列的概率推理意图识别技术结合上述两类技术特点，将概率统计建模方法引入机器学习算法模型中，人工建模方式保证模型本身规则的可解释性，同时针对模糊输入特征采用黑盒子方法快速完成推理过程。然而，由于此类技术基于人工建模形式，模型复杂度势必有限，故适合推理手眼输入等浅层次的交互意图，对于推理认知模型等深层次交互意图则难以被描述。常用的基于时间序列概率推理意图识别模型有：隐马尔可夫模型（Hidden Markov Models, HMM）[76-77]、贝叶斯分类模型等。

4.3.1 隐马尔可夫模型

隐马尔可夫模型是一种关于时序的统计学生成模型，用来描述一个含有隐含未知参数的马尔科夫链生成不可观测的状态序列、再由状态序列生成观测序列的马尔可夫过程，如图 4-5。它从可观察的参数中确定马尔可夫过程的隐含参数，再基于这些参数来做模式识别等进一步分析。Deng 等[76]综述了 HMM 模型在驾驶意图识别领域的应用，其指出 HMM 模型优点在于具有动态处理数据和时间模式识别的能力。Coutrot 等[41]使用 HMM 模型进行眼动行为扫描路径的建模与分类，通过状态转移概率矩阵直观描述 HMM 模型下由当前状态向未来状态的转移概率。Lang 等[77]将 HMM 模型应用于人机协作系统，提出了面向老年人的意图识别流程。然而，HMM 模型无法直接利用原始数据，必须预先进行数据处理，并且在训练之前必须知道隐藏状态的分类数量，因此该算法不适合长期预测系统

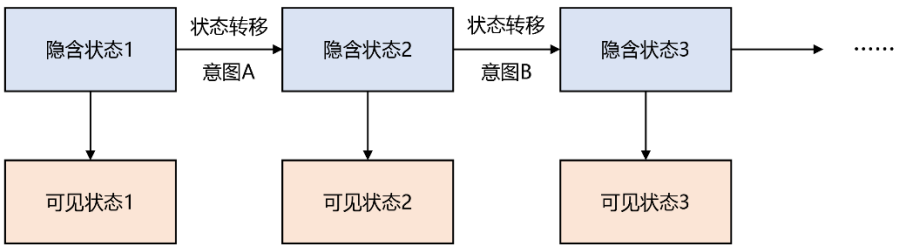

图 4-5 隐马尔克夫过程简化模型

4.3.2 贝叶斯分类模型

贝叶斯分类模型是以贝叶斯理论为基础的统计学生成模型。在统计推理问题中，极大似然估计和贝叶斯估计是最常用的参数估计方法，分别代表着频率学派和贝叶斯派的观点[78]。前者认为概率模型中参数是固定的未知常数，应将样本分布情况作为事件出现概率，其假设条件非常严格，要求训练集要符合真实数据分布；后者认为参数随机而样本固定，参数的取值将会影响概率分布，其优点在于基于最小风险论，根据后验概率最大者作为预测结果错误率低，且能够维持先验和观测对象间的平衡。

基于贝叶斯分类模型的分类算法涉及贝叶斯网络（Bayesian Neural Network, BNN）、朴素贝叶斯模型（Naive Bayes Classifier, NBC）与动态贝叶斯网络（Dynamic Bayesian Network, DBN），其中最常用的算法为动态贝叶斯网络。贝叶斯网络是借助有向无环图（Directed Acyclic Graph, DAG）来刻画属性之间的依赖关系，并采用条件概率表来描述属性的联合概率分布[79]，而动态贝叶斯网络在其基础上，引入时间概念，建立临近时刻随机变量间的依赖关系。时间特征的加入使得动态贝叶斯网络能够通过累积先验信息从而降低不同层次下推理过程中的不确定性，从而显著提升推理过程的准确度。

目前，动态贝叶斯网络已经广泛应用于小样本的意图识别与意图推理问题中。Yi 等[81]将 DBN 应用于手眼协同任务，基于眼睛与手的数据以及场景状态，动态贝叶斯网络可有效识别出三明治制作步骤中的各项任务意图。易鑫等[46]将动态贝叶斯网络应用于虚拟现实中的空中打字任务，基于手指位姿信息实现对用户文字输入意图的较为准确的解释与预测。李秀明[78]将基于时间转移概率的动态贝叶斯网络应用于战场作战场景，通过分析敌机飞行高度、雷达开关、速度和俯仰角等状态特征完成对敌方目标的战术意图推理。然而，动态贝叶斯网络仍然存在你需要复杂的网络定义，且在高维输入和过大样本等使用情形中表现不佳等局限性。

第五章 上下文场景中的多模态意图识别应用领域

基于上述意图特征提取与意图识别算法，需对当前多模态意图识别技术进行实践验证，评估其在具体应用场景的效果。本章将整理基于上下文场景的多模态意图识别在人机交互、战场作战、自然驾驶、日常任务、人机协作等领域的实际应用情况，并将本文出现的所有应用案例依据应用领域、应用场景、意图输入模态、意图特征、意图识别方法、意图识别结果整理在附录 I 中。

5.1 人机交互中的意图识别应用

人机交互中的意图是用户对计算机的互动期望，其意图识别聚焦于人机系统各交互模态的意图特征提取方法和意图分类技术，且特别侧重于觉察上下文的隐式人机交互，强调用户能透明地使用计算机系统，进而将精力集中于任务或情境的互动上，而非集中于和计算机系统的交互上[7]。其中，意图相关的人机交互方式设计和用户界面设计是人机交互中的意图识别的主要应用方向。在二维人机交互领域，Karaman 等[4]探究基于笔的交互意图识别，针对手绘笔和电子绘图板互动过程中按钮式快捷键使用不便的问题，提出通过 SVM 模型分析眼动特征与笔的输入特征，实现计算机对用户快捷键意图的自动识别。此外，Karaman[20]根据其提出的笔眼意图识别方式，设计了不同层次用户意图反馈的界面布局方案，并对界面可用性展开评估。在三维人机交互领域，易鑫[46]提出一种虚拟现实的空中裸手打字方式，通过分析用户手势关节与触点位置预测出用户想要输入的字母。

5.2 战场作战中的意图识别应用

战场作战中的意图聚焦于飞行员对战场态势的感知和战术指令目标。一方面，战场作战目标的动态属性与战场环境随着时间变化而变化，且作战行动具有一定的隐蔽性和欺骗性，导致战场作战中的意图识别难度加大；另一方面，战场作战意图知识表示难度较大，需要对军事专家的经验知识进行明确的组织、抽象和描述。建立准确的战场作战意图描述和基于战场态势感知的意图识别技术是该领域的主要应用方向。王崴等[82]融合脑电和眼动模态的意图特征，基于 SVM-DS 模

型完成飞行员战场操作意图的分类。Teng 等[75]基于相应作战环境定义敌方作战意图空间集，使用门控循环单元从敌机目标飞行参数中提取出敌方作战意图。

5.3 自然驾驶中的意图识别应用

自然驾驶中的意图是驾驶员对执行一系列未来车辆控制动作的态度，其意图识别聚焦于驾驶员的驾驶控制、驾驶接管、认知错误等方面意图识别以提高驾驶安全性。Martin 等[83]采集驾驶员注视次数、扫视时间、扫视频率等眼球运动相关特征并对自然驾驶行为中的眼动特征进行动力学建模，并基于上下文的扫描路径进行驾驶行为的变道意图预测。Yang 等[84]使用 Kinect 深度相机记录不同程度危险任务下的驾驶员头部与上半身的位姿特征，并通过前馈神经网络来预测当前驾驶员所处任务的危险程度。

5.4 日常任务中的意图识别应用

日常任务中的意图聚焦于人的日常活动，基于日常任务的意图识别已经在行人导航、视觉搜索、远程协作等日常活动中得到了广泛应用。Liao 等[72]开展一项真实世界的行人导航实验，记录用户在五种常见导航任务下的眼动特征，并采用随机森林预测用户当前导航任务。Halverson 等[85]基于 EPIC 认知模型建模，从认知心理学角度解释视觉搜索过程中眼动意图变化规律。

5.5 人机协作中的意图识别应用

在人机协作领域中，意图识别技术使机器人等智能体主动感知人的指令意图感知，提升人机协作效率。Wang 等[86]基于 HMM 模型提出基于机械臂的意图识别算法，机械臂通过分析在虚拟现实环境下人控制手柄的指向预测人的焊接意图。Lang 等[77]提出一种基于 HMM 模型的反向主动融合多模态意图的人机协同算法，通过分析用户的语音手势和姿势信息，完成任务协作意图的推理，并对该类意图识别算法下人机协作的信任程度给予有效评价。

第六章 总结与展望

6.1 总结

伴随着多模态传感技术、人工智能技术的飞速发展，人机关系正在发生着重大变革，人与计算机的互动不断由显式命令走向隐式感应、由单模态走向多模态、由瞬时状态的模式识别走向察觉上下文的行为理解。在人机融合的大趋势下，实现基于上下文场景的多模态意图识别是使计算机更贴合人的认知感知能力、建立由人自然但模糊的意图表达到计算机智能准确高效的感知响应间默契关系的必经阶段。总体来讲，目前上下文场景下的多模态意图识别的研究涉及到两个全然不同的领域——对“人”的认知感知机理探究领域和对“机”的机器学习模式识别等人工智能领域，这需要建立对“人”和“机”的双向理解。更进一步说，人机交互研究需要从人的认知感知、计算机的响应决策双向着手，建立人的高级感知认知加工和计算机的高效智能系统的强耦合关系，构成双向感知、双向理解的人机融合共生体。

6.2 展望

当前基于上下文场景的多模态意图识别相关研究仍处于起步阶段，未来研究可从如下方面开展更广泛的探索：

1.优化现有意图特征提取技术。由于目前硬件设备限制、特征提取算法的不准确性以及人生理模态特征的模糊性，现有意图特征提取技术仍然无法达到计算机意图识别算法的输入要求，且部分技术存在一定侵入性与场景限制。提取稳定、高质量的多模态意图特征，建立自由非侵入的意图特征提取方法是实现高效准确的基于上下文场景多模态意图识别的必要研究方向。

2.扩大意图识别技术适用范围。目前用户意图特征提取与意图识别方法高度特定于任务和情境，并且常基于实验室的受控环境下开展，难以在现实生活条件下投入应用。因此，应扩大意图识别技术的泛用度，注重更高复杂度的现实场景下的多模态意图识别应用。

3.建立完善的意图识别评价机制。基于上下文场景的多模态意图识别到目前

为止大多致力于客观层面意图识别准确率的衡量，而用户对意图识别技术的可用性感受、信任程度等评价机制仍未得到具体应用。因此，建立主客观结合的意图识别技术评价机制对于衡量人机融合程度与人机融合效果优劣具有重要意义。

参考文献

[1] TURK M. Multimodal interaction: A review[J]. Pattern recognition letters, 2014, 36: 189-195.

[2] PREMACK D, WOODRUFF G. Does the chimpanzee have a theory of mind?[J]. Behavioral and brain sciences, 1978, 1(4): 515-526.

[3] SHARMA R, PAVLOVIĆ V, HUANG T. Toward multimodal human–computer interface [M]. Advances in image processing and understanding: A Festschrift for Thomas S Huang, 2002: 349-365.

[4] KARAMAN Ç, SEZGIN T. Gaze-based prediction of pen-based virtual interaction tasks[J]. International Journal of Human-Computer Studies, 2015, 73: 91-106.

[5] SHI Y. Interpreting user input intention in natural human computer interaction[C]//Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. 2018: 277-278.

[6] SCHMIDT A. Implicit human computer interaction through context[J]. Personal technologies, 2000, 4(2): 191-199.

[7] 徐光祐,陶霖密,史元春,等.普适计算模式下的人机交互[J].计算机学报,2007(07):1041-1053.

[8] MORAN T, DOURISH P. Introduction to this special issue on context-aware computing[J]. Human–Computer Interaction, 2001, 16(2-4): 87-95.

[9] HUANG C, ANDRIST S, SAUPPÉ A, et al. Using gaze patterns to predict task intent in collaboration[J]. Frontiers in psychology, 2015, 6: 1049.

[10] ZHOU T, CHEN M, WANG Y, et al. Information entropy-based intention prediction of aerial targets under uncertain and incomplete information[J]. Entropy, 2020, 22(3): 279.

[11] XING Y, LV C, WANG H, et al. Driver lane change intention inference for intelligent vehicles: framework, survey, and challenges[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4377-4390.

[12] KELLEY R, TAVAKKOLI A, KING C, et al. Context-based bayesian intent recognition[J]. IEEE Transactions on Autonomous Mental Development, 2012, 4(3): 215-225.

[13] ÇIĞ Ç, SEZGIN T. Gaze-based prediction of pen-based virtual interaction tasks[J]. International Journal of Human-Computer Studies, 2015, 73: 91-106.

[14] MALLE B, KNOBE J. The folk concept of intentionality[J]. Journal of experimental social psychology, 1997, 33(2): 101-121.

[15] ASTINGTON J W. The child's discovery of the mind[M]. Harvard University Press, 1993.

[16] BLAKEMORE S, DECETY J. From the perception of action to the understanding of intention[J]. Nature reviews neuroscience, 2001, 2(8): 561-567.

[17] D'ANDRADE R. A folk model of the mind[J]. Cultural models in language and thought, 1987: 112-148.

[18] 关志伟. 面向用户意图的智能人机交互[D]. 中国科学院软件研究所, 2001.

[19] LUKANDER K, TOIVANEN M, PUOLAMÄKI K. Inferring intent and action from gaze in naturalistic behavior: a review[J]. International Journal of Mobile Human Computer Interaction (IJMHCI), 2017, 9(4): 41-57.

[20] KARAMAN Ç, SEZGIN T. Gaze-based predictive user interfaces: Visualizing user intentions in the presence of uncertainty[J]. International Journal of Human-Computer Studies, 2018, 111: 78-91.

[21] 马丽莎,吕健,潘伟杰,单军军,平正强.基于眼动模式的隐式意图识别分类方法研究[J].图学学报,2017,38(03):332-340.

[22] O'CONNELL T, WALTHER D. Dissociation of salience-driven and content-driven spatial attention to scene category with predictive decoding of gaze patterns[J]. Journal of vision, 2015, 15(5): 20-20.

[23] JERALD J. The VR book: Human-centered design for virtual reality[M]. Morgan & Claypool, 2015.

[24] LAVIOLA JR, KRUIJFF E, MCMAHAN R, et al. 3D user interfaces: theory and practice[M]. Addison-Wesley Professional, 2017.

[25] SHERMAN W, CRAIG A. Understanding virtual reality: Interface, application, and design[M]. Morgan Kaufmann, 2018.

[26] BETTINI C, BRDICZKA O, HENRICKSEN K, et al. A survey of context modelling and reasoning techniques[J]. Pervasive and mobile computing, 2010, 6(2): 161-180.

[27] WEI D, CHEN L, ZHAO L, et al. A Vision-Based Measure of Environmental Effects on Inferring Human Intention During Human Robot Interaction[J]. IEEE Sensors Journal, 2021,

22(5): 4246-4256.

[28] KOOCHAKI F, NAJAFIZADEH L. Eye gaze-based early intent prediction utilizing CNN-LSTM[C]//2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2019: 1310-1313.

[29] XUE J, LAI K. Dynamic gripping force estimation and reconstruction in EMG-based human-machine interaction[J]. Biomedical Signal Processing and Control, 2023, 80: 104216.

[30] CHENG Y, TOMIZUKA M. Long-Term Trajectory Prediction of the Human Hand and Duration Estimation of the Human Action[J]. IEEE Robotics and Automation Letters, 2021, 7(1): 247-254.

[31] THYE M, MURDAUGH D, KANA R. Brain mechanisms underlying reading the mind from eyes, voice, and actions[J]. Neuroscience, 2018, 374: 172-186.

[32] LI X, SHENG Q, PANG C, et al. Effective approaches in human action recognition[C]//2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2013: 1-7.

[33] 葛列众.工程心理学[M].上海:华东师范大学出版社,2017:163-164.

[34] CORBETTA M, SHULMAN G. Control of goal-directed and stimulus-driven attention in the brain[J]. Nature reviews neuroscience, 2002, 3(3): 201-215.

[35] JUST M, CARPENTER P. A theory of reading: from eye fixations to comprehension[J]. Psychological review, 1980, 87(4): 329.

[36] CASTIELLO U. Understanding other people's actions: intention and attention[J]. Journal of Experimental Psychology: Human Perception and Performance, 2003, 29(2): 416.

[37] YARBUS A. Eye movements and vision[M]. Springer, 2013.

[38] JACOB R. The use of eye movements in human-computer interaction techniques: what you look at is what you get[J]. ACM Transactions on Information Systems (TOIS), 1991, 9(2): 152-169.

[39] CHEN X, HOU W. Gaze-Based Interaction Intention Recognition in Virtual Reality[J]. Electronics, 2022, 11(10): 1647.

[40] KAR A, CORCORAN P. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms[J]. IEEE Access, 2017, 5: 16495-

16519.

[41] COUTROT A, HSIAO J, CHAN A. Scanpath modeling and classification with hidden Markov models[J]. Behavior research methods, 2018, 50(1): 362-379.

[42] LE MEUR , LIU Z. Saccadic model of eye movements for free-viewing condition[J]. Vision research, 2015, 116: 152-164.

[43] BEDNARIK R, VRZAKOVA H, HRADIS M. What do you want to do next: a novel approach for intent prediction in gaze-based interaction[C]//Proceedings of the symposium on eye tracking research and applications. 2012: 83-90.

[44] JANG Y, MALLIPEDDI R, LEE M. Identification of human implicit visual search intention based on eye movement and pupillary analysis[J]. User Modeling and User-Adapted Interaction, 2014, 24(4): 315-344.

[45] YASEN M, JUSOH S. A systematic review on hand gesture recognition techniques, challenges and applications[J]. PeerJ Computer Science, 2019, 5: e218.

[46] YI X, CHEN Y, SHI Y. Bayesian method for intention prediction in pervasive computing environments[J]. Scientia Sinica (Informationis), 2018.

[47] KHAN R, IBRAHEEM N. Hand gesture recognition: a literature review[J]. International journal of artificial Intelligence & Applications, 2012, 3(4): 161.

[48] SAGAYAM K, HEMANTH D. ABC algorithm based optimization of 1-D hidden Markov model for hand gesture recognition applications[J]. Computers in Industry, 2018, 99: 313-323.

[49] NOVAK D, OMLIN X, LEINS-HESS R, et al. Predicting targets of human reaching motions using different sensing technologies[J]. IEEE Transactions on Biomedical Engineering, 2013, 60(9): 2645-2654.

[50] XUE Y, JU Z, XIANG K, et al. Multimodal human hand motion sensing and analysis—A review[J]. IEEE Transactions on Cognitive and Developmental Systems, 2018, 11(2): 162-175.

[51] WANG W, LI R, DIEKEL Z, et al. Controlling object hand-over in human–robot collaboration via natural wearable sensing[J]. IEEE Transactions on Human-Machine Systems, 2018, 49(1): 59-71.

[52] GE S, DING M Y, ZHANG Z, et al. Temporal-spatial features of intention understanding based on EEG-fNIRS bimodal measurement[J]. IEEE Access, 2017, 5: 14245-14258.

[53] FELEKE A, BI L, FEI W. EMG-based 3D hand motor intention prediction for information transfer from human to robot[J]. Sensors, 2021, 21(4): 1316.

[54] BELKHIRIA C, BOUDIR A, HURTER C, et al. EOG-Based Human–Computer Interface: 2000–2020 Review[J]. Sensors, 2022, 22(13): 4914.

[55] WOLPAW J. Brain-computer interfaces (BCIs) for communication and control[C]//Proceeding of the 9th international ACM SIGACCESS conference on Computers and accessibility. 2007: 1-2.

[56] FARINA D, MERLETTI R, ENOKA R. The extraction of neural strategies from the surface EMG[J]. Journal of applied physiology, 2004, 96(4): 1486-1495.

[57] XI X, MA C, YUAN C, et al. Enhanced EEG–EMG coherence analysis based on hand movements[J]. Biomedical Signal Processing and Control, 2020, 56: 101727.

[58] XI X, MA C, YUAN C, et al. Enhanced EEG–EMG coherence analysis based on hand movements[J]. Biomedical Signal Processing and Control, 2020, 56: 101727.

[59] XIE H, GUO T, BAI S, et al. Hybrid soft computing systems for electromyographic signals analysis: a review[J]. Biomedical engineering online, 2014, 13(1): 1-19.

[60] NI P, LI Y, LI G, et al. Natural language understanding approaches based on joint task of intent detection and slot filling for IoT voice interaction[J]. Neural Computing and Applications, 2020, 32(20): 16149-16166.

[61] LIU B, LANE I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv preprint arXiv:1609.01454, 2016.

[62] CID F, MORENO J, BUSTOS P, et al. Muecas: A multi-sensor robotic head for affective human robot interaction and imitation[J]. Sensors, 2014, 14(5): 7711-7737.

[63] RIBONI D, BETTINI C. COSAR: hybrid reasoning for context-aware activity recognition[J]. Personal and Ubiquitous Computing, 2011, 15(3): 271-289.

[64] DE STEFANI E, DE MARCO D. Language, gesture, and emotional communication: An embodied view of social interaction[J]. Frontiers in Psychology, 2019, 10: 2063.

[65] BLAKE R, SHIFFRAR M. Perception of human motion[J]. Annual review of psychology, 2007, 58: 47.

[66] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal

analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125.

[67] TAPUS A, BANDERA A, VAZQUEZ-MARTIN R, et al. Perceiving the person and their interactions with the others for social robotics–a review[J]. Pattern Recognition Letters, 2019, 118: 3-13.

[68] WU S, ZHOU L, HU Z, et al. Hierarchical Context-Based Emotion Recognition With Scene Graphs[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[69] ELAKKIYA R, VANITHA V. Interactive real time fuzzy class level gesture similarity measure based sign language recognition using artificial neural networks[J]. Journal of Intelligent & Fuzzy Systems, 2019, 37(5): 6855-6864

[70] ZHOU T, CHEN M, WANG Y, et al. Information entropy-based intention prediction of aerial targets under uncertain and incomplete information[J]. Entropy, 2020, 22(3): 279.

[71] BOISVERT J, BRUCE N. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features[J]. Neurocomputing, 2016, 207: 653-668.

[72] LIAO H, DONG W, HUANG H, et al. Inferring user tasks in pedestrian navigation from eye movement data in real-world environments[J]. International Journal of Geographical Information Science, 2019, 33(4): 739-763.

[73] KOOCHAKI F, NAJAFIZADEH L. Predicting intention through eye gaze patterns[C]//2018 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2018: 1-4.

[74] WANG S, ZHAO X, YU Q, et al. Identification of driver braking intention based on long short-term memory (LSTM) network[J]. IEEE Access, 2020, 8: 180422-180432.

[75] TENG F, SONG Y, WANG G, et al. A GRU-based method for predicting intention of aerial targets[J]. Computational Intelligence and Neuroscience, 2021, 2021.

[76] DENG Q, SÖFFKER D. A Review of the current HMM-based Approaches of Driving Behaviors Recognition and Prediction[J]. IEEE Transactions on Intelligent Vehicles, 2021.

[77] LANG X, FENG Z, YANG X, et al. HMMCF: A human-computer collaboration algorithm based on multimodal intention of reverse active fusion[J]. International Journal of Human-Computer Studies, 2023, 169: 102916.

[78] 李秀明.基于贝叶斯推理的目标意图识别方法研究[D].哈尔滨工程大学,2020.

[79] GRIFFITHS T, KEMP C, TENENBAUM J. Bayesian models of cognition[J]. 2008.

[80] HAYHOE M, BALLARD D. Modeling task control of eye movements[J]. Current Biology, 2014, 24(13): R622-R628.

[81] YI W, BALLARD D. Recognizing behavior in hand-eye coordination patterns[J]. International Journal of Humanoid Robotics, 2009, 6(03): 337-359.

[82] 王崴,赵敏睿,高虹霓,等.基于脑电和眼动信号的人机交互意图识别[J].航空学报, 2021,42(02):292-302.

[83] MARTIN S, VORA S, YUEN K, et al. Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction[J]. IEEE Transactions on Intelligent Vehicles, 2018, 3(2): 141-150.

[84] XING Y, LV C, ZHANG Z, et al. Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition[J]. IEEE Transactions on Computational Social Systems, 2017, 5(1): 95-108.

[85] HALVERSON T, HORNOF A. A computational model of “active vision” for visual search in human–computer interaction[J]. Human–Computer Interaction, 2011, 26(4): 285-314.

[86] WANG Q, JIAO W, YU R, et al. Virtual reality robot-assisted welding based on human intention recognition[J]. IEEE Transactions on Automation Science and Engineering, 2019, 17(2): 799-808.

附录 I 上下文场景中的多模态意图识别应用案例汇总

参考文献	应用领域	应用场景	输入模态	输入意图特征	意图识别技术	意图识别效果
Karaman[4]	人机交互	手写笔交互	眼动行为	注视位置 手写笔位置	SVM	88%准确率
Huang[9]	日常任务	三明治制作	眼动	注视数据	SVM	76%任务预测准确率
Zhou[10]	战场作战	空中目标意图预测	行为	空中目标飞行状态参数	LSTM 决策树	-
Kelley[12]	人机协作	机器人理解代理人意图	行为	人类行为 视频图像	HMM	100%任务预测准确率
关志伟[18]	人机交互	文字编辑	行为	笔尖位置	组合神经网络	70%准确率
Karaman[20]	人机交互	手写笔交互	-	-	-	意图反馈界面设计
马丽莎[21]	其他	产品设计	眼动	瞳孔大小 注视时间/次数	NN SVM	SVM 比 NN 预测准确率高 85%准确率
O'Connell[22]	日常任务	视觉搜索	眼动	AOI	-	-
Wei[27]	人机协作	行为识别	行为	人体骨骼点 环境图像	LSTM MLP	98.8%准确率
Koochaki[28]	日常任务	视觉搜索	眼动	AOI	CNN LSTM	82.27%准确率
Xue[29]	人机协作	机械臂控制	生理信号	EMG	CNN	离线 92.57% 在线 82.05%
Cheng[30]	人机协作	装配任务	行为	人体骨骼点 人体位姿	LSTM	估计人未来行动的持续时间

参考文献	应用领域	应用场景	输入模态	输入意图特征	意图识别技术	意图识别效果
Chen[39]	人机交互	虚拟现实交互场景	眼动	三维注视点	RF/GNDT LR/SVM	GNDT 效果最好
Coutrot[41]	日常任务	视觉搜索	眼动	AOI 扫描路径	HMM	55.9%准确率
Bednarik[43]	人机交互	注视交互意图	眼动	注视位置	SVM	76%准确率
Jang[44]	日常任务	视觉搜索	眼动	注视 AOI 瞳孔反应	SVM	90%准确率
Yi[46]	人机交互	空中打字	手势	手指关节点 手指落点	DBN	支持空中双手盲打输入
Sagayam[48]	人机交互	手势识别	手势	手指关节点	HMM	73.59%准确率
Novak[49]	人机交互	多模态交互	脑电 眼电 肌电 眼动 手势	ERP EOG EEG 注视位置 手部动作	有监督机器学习	探究不同模态对预测人体到达目标的性能
Wang[50]	人机协作	机械臂控制	生理手势	EMG 手部运动	HMM SVM	82.08%评价准确率
Feleke[53]	人机协作	机械臂控制	生理手势	EMG 手部位姿	RFNN	-
Riboni[63]	日常任务	活动识别	行为	人类行为 环境信息	有监督机器学习	89.2%准确率
Boisvert[71]	日常任务	视觉搜索任务识别	眼动	AOI	随机森林 注视速度	-

参考文献	应用领域	应用场景	输入模态	输入意图特征	意图识别技术	意图识别效果
Liao[72]	日常任务	现实场景导航任务	眼动	注视扫视眨眼瞳孔直径	随机森林	67%准确率
Koochaki[73]	日常任务	视觉搜索	眼动	AOI 扫描路径	SVM	95.7%准确率
Wang[74]	自然驾驶	制动意图	行为	踏板动作	LSTM	95%以上准确率
Teng[75]	战场作战	空中敌方目标意图	行为	敌机信息	GRU(类LSTM)	89.7%准确率
Lang[77]	人机协作	任务意图	手势行为	语音手势行为姿势行为	HMM	97.8%准确率
Yi[81]	日常任务	三明治制作	眼动手势	注视位置手势动作	DBN	-
王崴[82]	自然驾驶	决策任务	生理眼动	EEG 注视位置瞳孔直径注视时间眼跳幅度	SVM	平均准确率92.34%
Martin[83]	自然驾驶	变道意图	眼动	注视次数注视时间扫视频率	-	准确率 75%
Xing[84]	自然驾驶	识别驾驶任务	行为	人体位姿	随机森林	准确率 80%
Wang[86]	人机协作	机器人辅助焊接	手势	手势动作	HMM	提高机器人辅助焊接的效果