多模态交互中的目标选择技术

周小舟，宗承龙，郭一冰，贾乐松，杜晓茜，薛澄岐
（东南大学 机械工程学院，南京 210000）

摘要：目的 伴随着各类传感器、计算机识别算法与计算机网络的发展，人机交互已不再拘泥于应用键盘、鼠标等传统输入模态，而是进一步拓展为应用人的多种行为模态。**方法** 多模态交互可以捕获多个模态下人的显性行为信息，组合并分辨多个模态信息的不同形式，现已应用在虚拟现实、增强现实、混合现实、遥操作、普适交互等场景下，可以优化这些场景下的人机交互。在自然交互中，多点触摸、自然语音、手势体感、眼动追踪是常用的交互模态，展现出了较优的效果。而多模态交互技术按照选择、相继、并发和互补的方式组合两个及以上的输入模式，借助多种非侵入式的传感器，识别天然形成的人类语言和行为。该技术的核心是基于各种传感器所捕获的多模态信息来预测人的交互意图，提升机器对人的行为理解，这可以很好地应用到目标选择任务之中。**结论** 在人机交互系统中，目标选择任务是一种基础任务，已经有菲兹定律模型、优化初始脉冲模型和层叠效应模型等行为模型在多个模态交互下对其进行了描述，这些理论都有助于开发多模态交互技术。综上所述，对多模态交互的目标选择技术研究可以有效地推动人机交互向自然交互迈进。

关键词： 多模态交互；目标选择技术；自然交互；交互意图
中图分类号： TB472 **文献标识码：** A **文章编号：** 1001-3563(2022)04-0036-09
DOI： 10.19554/j.cnki.1001-3563.2022.04.004

A Review of Target Selection Techniques in Multimodal Interaction

ZHOU Xiao-zhou, *ZONG Cheng-long*, *GUO Yi-bing*, *JIA Le-song*, *DU Xiao-xi*, *XUE Cheng-qi*
(School of Mechanical Engineering, Southeast University, Nanjing 210000, China)

ABSTRACT: With the development of various sensors, computer recognition algorithms and computer networks, human-computer interaction (HCI) is no longer confined to the application of traditional input modes such as keyboards and mouse, but further expanded to the application of various modes driven by human behavior. Multimodal interaction can capture the explicit behavior information of multiple behavior modes and distinguish different forms of multimodal information after combination. Nowadays it has been applied in virtual reality, augmented reality, mixed reality, teleoperation, universal interaction and other settings to optimize human-computer interaction. Multi-touch, natural speech, gesture motion sensing and eye movement tracking are commonly used in natural interaction. And all of them show their advantages in specific fields. Multimodal interaction technology combines two or more input modes in selective, sequential, concurrent or complementary ways, using a variety of non-invasive sensors to identify naturally occurring human language and behavior. The core of this technology is to predict human interaction intention based on multi-modal information captured by various sensors and improve the understanding of machine to human behavior, which can be well applied to target selection tasks. In human-computer interaction system, target selection task is a basic task, which has been described by Fitz's law model, optimized initial pulse model and cascade effect model under multiple modal interaction. These theories are helpful for the development of multi-modal interaction technology. To sum up, the research on target selection technology of multimodal interaction can effectively push human-computer interaction forward to natural interaction.

KEY WORDS: multimodal interaction; target selection technology; natural interaction; interactive intention

收稿日期： 2021-09-10

作者简介： 周小舟（1982—），女，江苏人，博士，东南大学机械工程学院讲师，主要研究方向为先进交互设计、大数据可视化、产品与品牌识别设计。
通信作者： 薛澄岐（1961—），男，江苏人，博士，东南大学机械工程学院教授，主要研究方向为工业设计。

随着各类传感器、计算机识别算法和计算机网络的发展，除了键鼠操作等交互方式以外，语音、手势、体感、眼动追踪等多个交互模态逐渐应用到了人机交互系统中。人机交互系统是包含软件、硬件以及使用者，连接人和计算机的系统[1]。在人机交互系统的发展过程中，人们开始探索用更加符合人的本源性认知和行为习惯的交互方式与计算机进行沟通，而人的意图表达是多模态的，与此对应的多模态交互也应运而生。多模态交互的核心是使计算机具有类人的感知功能，人机交互在一定程度上靠近人人交互的自然水平[2]。多模态交互要求计算机能够识别多种类型传感器所捕获的人的多模态交互行为，将其解码并生成计算机下一步动作的指令。当交互行为更符合人的本源性自然表达时，用户花费较少的认知资源便可以获取和传达信息，以达到提升人机交互效率和交互体验的目的[3]。近年来，多模态交互在虚拟现实[4]、增强现实[5]、混合现实[6]、遥操作[7]、普适交互[8]等场景下都有着广泛的应用前景。

1　交互模态

多模态（Multi-modal）的概念最早出现在语言学领域，延伸到社会符号学、教育学等多个领域[9]。在人机交互领域中，多模态也有着多种不同的解释，包括不同的信息感知通道、不同的信息呈现方式等。本文中的多模态是从计算机信息输入通道的角度上来阐述的，人机交互的模态包含传统的输入工具，常用的鼠标、键盘，以及轨迹球、摇杆等，以及较为新颖的输入模态，包括语音、多点触控、手势、体感、眼动追踪等。每个模态都有其独有的交互特点，包括设备优势与限制、交互能达到的精确度、稳定性，以及对用户造成的肌肉疲劳与认知负荷等。考虑到人类的交互行为是多个感受和执行通道共同作用的，因而将符合人行为学的交互模态融合起来作为人机交互的输入和输出方式，多模态交互更有利于人对计算机环境的感知和计算机对人意图的理解[10]。本文将对能提高交互自然性的输入模态，包括自然交互中常用的触摸、语音、手势体感、眼动追踪等作简要介绍。

1.1　触摸交互

触摸交互是指通过和触控屏幕接触而产生的一种二自由度的交互手势，根据交互手势中触摸点的位置、触摸状态和触摸点相对位移等特征转化为控制信号[11]。触控输入除了与显示屏直接接触的输入方式外，还包括笔尖和“悬空”的输入方式，本段主要介绍手指直接触摸的交互方式。触摸手势的属性可以分为触摸点属性和移动特征属性。触摸点属性包括与屏幕产生接触的点的个数及接触类型，现有的触控技术可以利用接触面积、压感、电阻等信息分辨手指或骨节等接触信息，并支持多点触控，允许用户利用双手或多手指进行触摸交互[12]。移动特征属性主要有点击、拖动、滑动、横扫、双击、放大、缩小、长按、旋转等[13]。根据手势的自然运动属性，手势可以表达很多含义，多点触控手势在功能上可以分为点按与辅助查找、滚动与缩放控制、全局控制三种类型[14]。

触摸交互技术作为一种提升自然性的输入方式，可以利用振动、和触感等输出反馈方式来提升交互的触觉体验[15]，其主要限制是对空间要求高，交互对象必须在用户的身体可达域范围内，在虚拟现实环境中交互绩效受限[16]。

1.2　语音交互

语音交互通常是指利用声音来实现信息的输入、输出、反馈及响应，是一种可以直接反映人类内心意图的人机交互方式，能达到以交谈式为核心的智能人机交互体验[10]。语音交互可以解放用户的双手，或是可以在双手已经被占用的前提下实现较大词汇量的交互功能，因而语音控制常被用于作为选择任务的补充方案。用户输出的语音经由语音识别、自然语言理解、对话管理、响应生成后，系统将对人的输入信息做出对应的交互响应[17]。借助语音交互技术，用户的操作可以穿透多重视觉层次，无视中间应用、网页和复杂环境等，实现直达用户想要的操作的交互目的[10]。

作为用户日常生活中熟悉的交互通道之一，利用语音实现人机交互可以减轻用户对交互任务的学习量，适用于非图形的命令交互与控制交互[10]，能够实现较为复杂的指令功能，有效减轻用户的肌肉疲劳度，并提升交互的自由度。然而这种交互方式也存在其限制。由于人的语言天然的具有模糊的特点，语音控制系统往往需要根据背景推理用户所表达的含义，在不确定语义背景的情况下，具有很高的错误率[18]；多用户之间的语音会产生干扰，语音输入的私密性也无法保障[19]。这些特性使得语音控制不常作为首选的输入方式，在应用场景的广度上受限。

1.3　手势与体感交互

手势与体感交互是指通过用户手部或肢体的静态姿势或动态动作来进行计算机指令输入，从而实现相关功能的交互方式。常用的手势与体感包括手部姿势与动作、臂部姿势与动作、头部姿势等[10]。一般来说，手势与体感动作通过计算机视觉或者穿戴式传感器跟踪的方法被设备跟踪并捕捉，从而作为指令的发送方向计算机传递发出指令的信号[20]。

与传统的键鼠交互方式相比，手势与体感交互是更具自然性的交互方式。在空间上，手势与体感交互打破了键鼠等交互方式中设备对用户的桎梏，用户可以在较远的距离上脱离实体来进行交互操作。人的手部动作有着丰富的可能性，作为交互输入方式具有自然性、灵活性、便捷性等优势[21]。尤其在虚拟现实环境中，它是能提升用户沉浸感的重要自然交互方式之

一。利用手和手臂姿势变化，用户可以完成诸如浏览网页、翻看书籍、放大缩小物体等交互动作。然而这种交互方式也仍存在一些亟待解决的交互缺陷。可用于指令的符合自然性的交互手势与体感有限，人们很难通过手势与体感动作完成大量且复杂的诸如文字输入这样的操作[22]。用户的指令动作和无意识的自然动作在识别过程中容易混淆，产生对指令起止判断等歧义，而造成弥达斯接触问题[23]。除此之外，手势与体感交互还具有响应时延和占用较多记忆资源等局限性[21]。

1.4 眼控交互

眼控交互是指通过对所获取的视线移动的位置、轨迹、速度、驻留时间等信息进行特征定义，将其作为计算机交互指令的交互方式[24]。按照交互主动性，眼控交互可分为基于视线的交互和视线辅助的交互两种[25]。将基于视线的交互作为独立的交互控制模态，容易造成视觉疲劳，因而多用在医疗、残疾人辅助设备等特殊场合。

按照眼动信息特征，主动的眼控方法可以分为凝视交互、眨眼交互、平滑追踪和眼势交互四种[26]。这四种眼控方式有着不同的交互逻辑。其中，凝视交互与传统鼠标交互逻辑相似，又因为其操作简单所以是目前应用最广的眼控交互方式[27]，然而其存在着不自然、费力等问题，还因为眼球的无意识抖动行为存在一定的精度问题[28]；眨眼交互对眼动追踪设备的时空分辨率要求最低，但可以使用的交互命令较少，且有意识眨眼与无意识眨眼在区分上有难度；平滑追踪的交互方式依赖于动态刺激[29]，其速度会影响平滑追踪的执行效果[30]；眼势交互可以在一定程度上规避弥达斯接触问题，但容易造成疲劳且学习成本较高。因此，目前通用型人机交互领域常以视线辅助的交互形式来有效地辅助其他交互模态实现人机交互行为。

2 目标选择任务中的行为模型

目标选择是人机交互的基础要素，本质是一种对用户交互意图的提取。在传统键鼠交互中，输入设备的输入信息是确定性的，物理设备的运动和光标之间存在明确的对应关系。而在强调交互自然性的多模态交互中，由于应用了人天然的输出模态，如空中手势、凝视、语言等均具有模糊性，输出信息与计算机指令的对应关系会变得模糊。为了达到自然交互的目的，就需要在多种模糊的模态中挖掘人的行为中确定的交互意图，而若要发现这种确定性，就必须建立明确的行为模型。构建指向选择任务的行为模型往往需要将任务划分成几个子阶段，多模态交互的优势之一就是可以给不同模态分配不同的子阶段任务，以此避免过度使用单模态造成的疲劳和单模态的技术缺陷[31]。

指向选择任务可以分为指向任务和选择任务两个子任务，对应着目标获取和验证确认的交互目的。为了简化模型，多模交互的发展初期通常在指向和选择（或操作）阶段各应用一个模态。常见的多模态技术多应用视线完成指向，应用手势进行选择或操作。例如，在虚拟现实中注视目标后用“捏”的动作移动目标[32]或用“抓握”的手势“握住”物体并移动[33]，在触摸屏交互中注视目标后点击屏幕任意位置对平面目标进行缩放旋转等操作[14]。由于人在目标操作尤其是高精度操作时具有注视目标的行为倾向[34]，因此这种多模交互方式具有提升交互绩效的实用价值。然而这种方法因为在指向阶段仅视线一种输入模式，对视线捕捉的精度要求较高，在目标较小或者完成精度要求高的应用场合容易产生交互失误而造成用户的挫折感和疲劳体验。为了更精确研究指向任务，Woodworth 等人率先提出了指向任务两阶段理论，将指向动作划分成快速弹射运动阶段和调整阶段[35]，该理论被后续研究者广泛应用，通过将指向过程分阶段分析和应用来获取更准确的指向数据和优化模型。目前，已有多个指向选择任务的行为理论可以用来指导多模态交互技术开发，包括菲兹定律、优化脉冲模型、层叠效应理论等。

2.1 菲兹定律

菲兹定律[36]是表达指向选择任务中用户完成任务所用时间的理论，是在人机交互领域少有的定量表达人机交互系统效果的理论模型。从信息论的观点来看，人输入到计算机的信息容量 C（比特/秒）取决于通信信道的带宽 B（s^{-1} Hz）、信号功率 S 和噪声功率 N，其关系如公式(1)所示：

$$C = B\log_2\left(\frac{S}{N}+1\right) \tag{1}$$

比照信息论的公式，MacKenzie[37]提出了目前被广泛采用的菲兹定律计算方式，他把完成选择任务所需要的时间 T 与目标的宽度 W 以及与目标的距离 A 建立了联系，并用其比例对数的线形回归模型来预测运动时间，如公式(2)所示：

$$T = a + b\log_2\left(\frac{A}{W}+1\right) \tag{2}$$

其中 a 与 b 是该回归方程的回归系数，而对数项则被称为难度系数 ID。系数 a 会受到确认动作等附加因素的影响，而 $1/b$ 则可以反映交互系统的性能，该性能通常称为吞吐量。

菲兹定律在一维到三维中都有应用。Wingrave 和 Bowman 的研究[38]表明，菲兹定律在虚拟三维环境中依然有效。在三维环境下，物体的 W 需要以其出现在用户视野里的视觉大小来表示，而 A 则需要进一步考虑用户手部的旋转角度。Poupyrev[39]则进一步将物体的 W 以物体出现在用户视野里的竖直与水平的角度进行定义。菲兹模型在三维物体的选择中得到

图 1　优化初始脉冲模型

Fig.1　Optimized initial impulse model

优化[40-42]。菲兹定律可以体现出人的指向选择任务的行为特征，研究表明除了各种以手为基础的交互之外，脚、头、眼睛的选择指向仍然满足菲兹定律[43]，因此菲兹定律可以作为多模态交互的一般性行为模型。

2.2　优化初始脉冲模型

基于指向任务两阶段理论，Meyer 等人提出的优化脉冲模型[44]常被用来解释用户执行选择任务时的手部运动。不同于菲兹定律的宏观预测和评估理念，该模型对任务过程做了更细致的描述，它将选择任务中的手部运动阶段区分成低精度快速移动的快速弹射运动阶段与高精度慢速移动的慢速调整阶段，用于描述在选择任务中不同阶段用户进行操作的速度与任务要求的变化。优化初始脉冲模型见图 1，慢速调整阶段出现在快速弹射运动阶段之后，这两个阶段使得人在执行此类交互动作时可以兼顾速度与精度。

人的生理特性导致人的肢体行为无法同时兼顾快速和精确的运动要求。一般情况下，用户所需选择的目标是随机分布在某个区域内的，这导致人的肌肉群必须做更微小的调节才可以完成选择[45]，而参与大范围快速移动的肌肉群往往较大，无法在兼顾速度的同时完成精确的选择。对于需要进行精确操作的慢速调整阶段，小肌肉群无法实现大范围的移动，但它们更加细分的可变性使其更容易完成细小的调整。

目前优化脉冲模型仍在不断的优化过程中。Piumsimboon 等人分阶段研究了弹射阶段校准阶段的输入模态[46]，分别测试了头和眼完成弹射阶段并与调整阶段其他模态结合的绩效和主观评价，证实了眼在速度上的优势和对设备及准确性要求带来的用户体验问题以及头指向的交互准确性和脖子疲劳问题等。邓成龙等人在两阶段理论的基础上基于远距离移动物体过程中对目标移动速度的观察，又将弹射阶段分为了加速阶段和减速阶段，建立了移动物体的三阶段理论[47]，该三阶段理论对指向任务中的普适性有待进一步的研究。还有研究表明，用户在选择任务中会自行平衡快速弹射运动阶段与慢速调整阶段[48]，这两个阶段并不是固定的且不可改善的。MacKenzie 等人进一步发现[49]，速度在时间序列上的变化取决于目标的宽度 W 以及与目标的距离 A，而不仅仅是难度系数 ID。A 会影响快速弹射运动阶段的最大速度，而 W 则影响慢速调整阶段所需要做的修正，这为借助该理论实现自适应交互提供了条件。

优化脉冲模型是对交互运动的细化，而人的多种行为模态都可以作为该模型中的运动指标来源，从而在菲兹定律的基础上进一步细化交互流程，对交互意图进行更详细地分析和定义，是实现无感的多模态交互的基础。

2.3　层叠效应理论

眼部运动的实时监测可以获取用户注意焦点、快速定位用户兴趣区，是多模态交互中意图捕获的基础。在眼睛的运动规律方面，Shimojo 等人提出了层叠效应理论[50]。在选择任务中，物体得到的注意越多，它被选择的概率就越大。该理论阐述了一个统计模型，在注意与决策之间搭建了桥梁，所包含的变量仅包含目标得到的注意。诸多神经行为学研究发现，当大脑在诸多刺激间进行选择时，人脑会首先对多个刺激赋值，随后再考虑应该选择哪个刺激[51]。表现在眼球运动上，在日常生活中需要进行决策、选择的任务里，人眼会不停地交替注视多个刺激，以完成刺激赋值进程[52]。视觉的层叠效应理论反映的就是这种赋值过程。

由于眼动的注视信息可以很好地反映人的注意力特征[53]，因此层叠效应常应用于借用眼动信息预测用户决策，进而在交互全过程完成前提前预测交互意图[54]。研究表明，人的注意力特征与人脑对刺激的赋值过程可以互为因果，不仅刺激本身的特性可以吸引人的注视，更长的被注视时间也可以导致该刺激的被选择概率提升[55]。随着决策过程的推进，这种双向促进的过程使得眼动特征与人的决策可以深度绑定，进而呈现出更加确定的结果。Smith 等人[56]也进一步研究了这两种效应的强弱，进一步发现其相对强弱在不同场景下有所不同。为了完成从眼动信息到决策信息与交互意图的预测，已经有很多研究者通过建立模型对层叠效应进行量化，在实验室环境下通过模型计算决策结果[57-58]。而在人机交互技术的应用领域，可以

使用神经网络完成眼动交互意图的识别和预测[59-60]。层叠效应理论所展示的是人的注意选择规律，便捷的眼动注视目标的获取设备和技术使得该理论具有广泛的应用前景，可以作为交互意图捕捉方式和多种交互模态共同实现更为快速精确和确定性的交互目的。

3 目标选择中的多模态融合方式

多模态交互技术是一种以协调的方式处理两个或多个输入模式，借助多种非侵入式的传感器，识别天然形成的人类语言和行为，以获取人的交互意图并输入计算机的技术[61]。由于传感器输入信息的组合，输入信息容量更高，所以具有超越单模态的输入效率。同时多个传感器输入信息可以相互作用，降低信息中的不确定性，多模态交互识别系统的准确率远高于单模态的输入。多模态交互技术具有比单模态交互技术更好地理解人的交互意图的理论基础。

人在进行意图表达时会自然地同时调用多个输出模态。例如人在指向目标物时，会转向、注视目标并用手指向目标；阐述复杂概念时，人会在语言表达的同时辅助空中手势的表达。因此，多模态交互技术是以本源性自然表达为目标的自然交互技术发展的必然趋势。由于交互情境的多样性和交互模态的适用性，多模态交互的模态融合方式具有多样性。剖析交互模态在融合方式上的特征，归纳了以下四种类型：选择型、相继型、并发型、互补型，多模态交互的模态融合方式见图 2。

选择型多模态交互，是指某一交互输入模态或组合均表示相同的语义信息，各模态输入信息在功能上都是等效的，用户自行选择或者根据场景自适应适配的交互融合方式。此类交互技术希望通过提供多种各具特点的输入模态，满足不同用户在不同场景下的偏好，提高用户输出意图的效率。在携带有语音助手的智能手机中，设置闹钟往往可以通过触摸或语音等不同的方式实现。一些研究也探索了模态的自动选择，以避免增加用户的认知负荷，例如 Pfeuffer[62]等人研究了人的注意力机制，并将注意力机制用于在手眼之间切换输入模态，借助这种自然的切换，更好地匹配了选择任务中所适合的模态，提高了输入效率。

相继型多模态交互，是指两个或多个输入模态在时间线上的不同时间段先后发挥作用，最终共同完成一个任务操作指令的交互融合方式。在此类系统中，前一种模态可以用于防止后一种通道错误的触发，并适时地激活后一种模态，为任务的不同阶段使用合适的输入模态，避免计算机错误地识别到了用户并不存在的交互意图。例如在多模态交互的一键通话界面中，语音模态从一个手势动作获得信息，并将语音输入激活。已经有学者采用这种组合方式来解决选择任务的精确度与速度问题。例如，MAGIC 指向技术使用头部信息初始化屏幕上的光标位置，之后再由鼠标接管光标[63]。Yang 提出了一种使用眼动进行粗略选择，使用触摸板进行精确选择的操控技术[64]。Koskinen 在外科手术领域开发了一种技术，他们通过提取手术刀上的注视点信息来确定画面的缩放幅度，以此来配合手执行不同精细度的手术操作[65]。在 Cordeiro 等人所开发的增强现实僵尸游戏中，面部识别所获取的头部朝向被用来完成游戏里射击动作的瞄准，触摸则被用于确认开火[66]。

并发型多模态交互，是指需要两个或两个以上的输入模态在同一时间段内触发才能完成一个任务的交互融合方式。其主要表现在时间段上的同步性，强

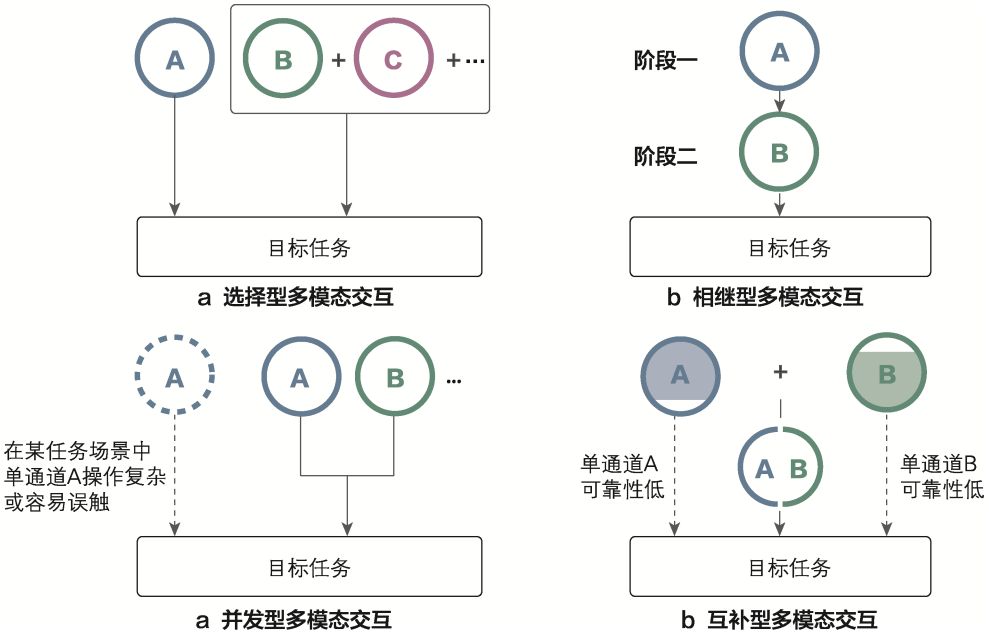

图 2 多模态交互的模态融合方式

Fig.2 Modality fusion methods in multimodal interaction

调不同交互模态需要在同一时间段内被调用，两种模态同时触发才能构成完整的语义。此类交互技术可降低单模态下的偶发启动，将多模态设定为彼此的互锁机制，提升交互操作的确定性。Pfeuffer 等人[32]所开发的技术就旨在通过这种方法消除眼动的弥达斯接触问题，该技术以眼动为指向，以一个捏合姿势为确认动作，只有当眼动选中目标且发出捏合的确认动作时，目标才被选中。这种组合方式也适用于涉及多维度信息的任务中，在为虚拟环境下某一物体赋予颜色的任务中，用户需要同时输入色彩和目标两个信息，EyeSeeThrough[67]技术让用户用手拿起一个调色板，当眼、调色板的颜色、目标共线时完成色彩的赋予。

互补型多模态交互，是指提取两个或多个输入模态的优势动作配合发出命令，共同完成一个任务，以消除交互意图中歧义的交互融合方式。设计互补型多模态交互任务的时候，需要针对每种交互模态的优势动作和交互响应来细化整个交互动作，以实现功能上的最优分配，考虑交互任务的具体实现场景来选择可以同时实现协同操作的交互模态来完成一系列交互动作。例如，Argelaguet 等人[41]采用眼的位置发出射线，并使用手腕转动引导射线的方向移动信息，同时规避了眼动的不稳定性以及手势射线容易被遮挡的问题。Bai 等人[68]在车内的选择场景下，通过手势进行选择指向，凝视信息用来确认指向的正确性，在不产生额外运动的情况下提升了选择的准确率。Li 等人[69]也在平板电脑上开发了相似的轻量化技术，以减少手势识别的误差。在 Sidenmark 等人[34]所开发的交互技术中，眼动信息与控制器信息相互补充，当控制器确定一个目标点时，该交互技术会隐式地对眼动仪进行校准，进而避免了用户频繁地主动校准眼动仪。

4　结语

伴随着各类传感器、计算机识别算法与计算机网络的发展，计算机对人的交互意图感知能力不断加强，多模态交互已成为人机交互的必然发展方向。多模态交互可以在对人交互行为分析的基础上实现非侵入、无感的自适应交互。而目标选择任务作为人机交互中的基础任务，具有任务典型性和研究必要性。当前多模态交互中的目标选择任务的优化方向包含以下四个方面。（1）建立基于意图捕捉的人机交互。目前的多模交互技术大多将选择任务划分成“指向+选择”或“指向+校正+选择”的分步形式，每一步之间需要用户通过手势或语音等方法明确告诉计算机步骤的切换，该行为极大地简化了计算机的工作却增加了用户的交互任务量。在未来的研究中，将这种用户的主动交互意图表达转化为计算机的主动交互意图识别，在计算机对人行为的充分理解的基础上建立基于人的意图捕捉的人机交互形式是多模交互中优化目标选择模型的必要研究方向。（2）多模态无缝融合交互技术。为了达到更自然的交互效果，已有很多研究都尝试了多模态交互，但是目前较多的多模态交互还停留在单模的组合上，即利用拼接单模的方式，让不同模态交互实现不同阶段的功能后再组合到一起：例如利用眼控完成指向后再用手势进行确认等。这种方法虽增加了操作的词汇量，有助于用户完成更多交互内容，却保留了单模的缺点，且不符合用户的自然交互习惯。因此实现多个模态间的无缝融合可以达到显著的优化模型的效果。（3）虚拟空间中交互的自然贴合度。虚拟现实、混合显示等虚拟空间为多模态交互提供了广泛的研究依托和应用场景。在虚拟现实中，交互过程与真实物理世界相似度的提高有利于提升用户的交互兴趣，而且能提升没有经验的使用者的交互能力[3]。当前的指向任务多采用单一的空间射线投射技术，然而用户的指向动作模型随目标距离等因素改变会产生变化，比如在指向远距离目标时目标的位置更接近于眼与手指间的延长线方向，指向近距离物体时更接近于手指指向方向[70]，因此可以借助多模态的方法提升指向选择任务模型与用户行为习惯的贴合度，从而提升交互自然性和准确性。（4）基于现实复杂场景的设计优化。大多数设计停留在单一的实验层面，缺乏实践应用，实验场景与实践场景差别大，仅停留在规律单一的实验场景以绩效评价证明设计方法的可用性。且前期对于技术的特点研究不充分，对于单模的缺陷认识停留在射线投射精度低、眼动数据不准确等，无法最大化发挥不同技术的优势、合理利用以降低单模态的缺陷造成的交互体验降低等问题，缺乏利用多模态实现复杂场景的交互案例。因此未来的研究应注重复杂、拟真场景下多模态技术的应用。

参考文献：

[1] CARD S K, MORAN T P, NEWELL A. The Psychology of Human-Computer Interaction[M]. University of Michigan Published: the Psychology of Human-computer Interaction, 2008.

[2] SIDENMARK L, GELLERSEN H. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality[J]. ACM Transactions on Computer-Human Interaction, 2019, 27(1): 1-40.

[3] MENDES D, FONSECA F, ARAUJO B, et al. Mid-air Interactions Above Stereoscopic Interactive Tables[C]. IEEE Symposium on 3D User Interfaces (3DUI), 2014.

[4] CASHION J, WINGRAVE C, JR J J L. Dense and Dynamic 3D Selection for Game-Based Virtual Environments[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(4): 634-642.

[5] CARMIGNIANI J, FURHT B, ANISETTI M, et al. Augmented Reality Technologies, Systems and Applications[J]. Multimedia Tools & Applications, 2011, 51(1):

341-377.

[6] MOHAMAD YAHYA FEKRI A, AJUNE WANIS I. A Review on Multimodal Interaction in Mixed Reality Environment[J]. IOP Conference Series: Materials Science and Engineering, 2019, 551(1): 8-12.

[7] STEINICKE F, ROPINSKI T, HINRICHS K. Multimodal Interaction Metaphors for Manipulation of Distant Objects in Immersive Virtual Environments[C]. The 13-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2005 in Co-operation with EUROGRAPHICS, 2005.

[8] SHOHEI Y, KOZO K, TAKEFUMI Y, et al. Construct Validity for Eye-hand Coordination Skill on a Virtual Reality Laparoscopic Surgical Simulator[J]. Surgical Endoscopy, 2019, 12(12): 2253.

[9] JAIMES A, SEBE N. Multimodal Human-computer Interaction:a Survey[J]. Computer Vision and Image Understanding, 2007, 108(1): 116-134.

[10] ZHANG F, DAI G, PENG X. A Survey on Human-computer Interaction in Virtual Reality[J]. SCIENTIA SINICA Informationis, 2016, 46(12): 1711-1736.

[11] TAKEOKA Y, MIYAKI T, REKIMOTO J. Z-touch: An Infrastructure for 3D Gesture Interaction in the Proximity of Tabletop Surfaces[J]. 2010(1): 91.

[12] SAE-BAE N, MEMON N, ISBISTER K, et al. Multitouch Gesture-Based Authentication[J]. IEEE Transactions on Information Forensics & Security, 2014, 9(4): 568-582.

[13] HEMANT SURALE. Barehand Mode Switching in Touch and Mid-Air Interfaces[GB/OL]. (2020-05-08) [2021-01-20]. http://hdl.handle.net/10012/15950.

[14] PFEUFFER K, ALEXANDER J, CHONG M K, et al. Gaze-touch: Combining Gaze with Multi-touch for Interaction on the Same Surface[C]. Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, 2014.

[15] BASER O, KONUKSEVEN E I. Kinematic Calibration of a 7 DoF Hapic Device[C]. Tallinn: 15th International Conference on Advanced Robotics (ICAR), 2011.

[16] LUBOS P, BRUDER G, ARIZA O, et al. Touching the Sphere: Leveraging Joint-centered Kinespheres for Spatial User Interaction[J]. SUI 2016 - Proceedings of the 2016 Symposium on Spatial User Interaction, 2016(1): 13-22.

[17] KASHEVNIK A, LASHKOV I, AXYONOV A, et al. Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin[J]. IEEE Access, 2021(9): 34986-35003.

[18] JUNG N, KIM G, SON CHUNG J. Spell My Name: Keyword Boosted Speech Recognition[J]. arXiv e-prints, 2021(1): 10.

[19] YUAN J H, LI-ZHOU A N, WANG H T. Research on the Teaching Application of Virtual Reality Technology in Equipment Construction Courses[J]. Sci-tech Innovation and Productivity, 2020(1): 10.

[20] SUN Y, WENG Y, LUO B, et al. Gesture Recognition Algorithm Based on Multi-scale Feature Fusion in RGB-D Images[J]. IET Image Processing, 2020, 14(15): 3662-3668.

[21] NIELSEN M, STORRING M, MOESLUND T B, et al. A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI[C]. Berlin, Heidelberg: International Gesture Workshop, 2003.

[22] MCAWEENEY E, ZHANG H, NEBELING M. User-driven Design Principles for Gesture Representations[C]. Canada: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.

[23] DREWS H. Eye Gaze Tracking for Human Computer Interaction[D]. lmu, 2010.

[24] CHUNG S E, Ryoo H Y. Functional/Semantic Gesture Design Factor Studies on Social Robot for User Experience Design[J]. International Journal of Smart Home, 2020, 14(1): 1-8.

[25] 李婷. 眼动交互界面设计与实例开发[D]. 杭州: 浙江大学, 2012.
LI Ting. Eye Movement Interactive Interface Design and Example Development[D]. Hangzhou: Zhejiang University, 2012.

[26] LAND M F, TATLER B W. Looking and Acting: Vision and Eye Movements in Natural Behaviour[M]. Looking and Acting: Vision and Eye Movements in Natural Behaviour, 2012.

[27] BATES R, DONEGAN M, ISTANCE H O, et al. Introducing COGAIN: Communication by Gaze Interaction[J]. Universal Access in the Information Society, 2007, 6(2): 159-166.

[28] VIDAL M, BULLING A, GELLERSEN H. Pursuits: Spontaneous Interaction with Displays Based on Smooth Pursuit Eye Movement and Moving Targets[C]. Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2013.

[29] NIU Y, LI X, YANG W, et al. Smooth Pursuit Study on An Eye-Control System for Continuous Variable Adjustment Tasks[J]. International Journal of Human-Computer Interaction, 2021(1): 1-11.

[30] 朱潇潇. 基于眼势的眼控界面交互设计研究[D]. 南京: 东南大学, 2019.
ZHU Xiao-xiao. Research on Eye Control Interface Interaction Design Based on Eye Potential[D]. Nanjing: Southeast University, 2019.

[31] DENG S. Multimodal Interactions in Virtual Environments Using Eye Tracking and Gesture Control[D]. Bournemouth: Bournemouth University, 2018.

[32] PFEUFFER K, MAYER B, MARDANBEGI D, et al. Gaze+ Pinch Interaction in Virtual Reality[C]. Proceedings of the 5th Symposium on Spatial User Interaction, 2017.

[33] RYU K, LEE J J, PARK J M. GG Interaction: A

Gaze-grasp Pose Interaction for 3D Virtual Object Selection[J]. Journal on Multimodal User Interfaces, 2019, 13(4): 383-393.

[34] Sidenmark L, Lundström A. Gaze Behaviour on Interacted Objects During Hand Interaction in Virtual Reality for Eye Tracking Calibration[C]. Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, 2019.

[35] WOODWORTH R S. The Accuracy of Voluntary Movement[J]. Psychological Monographs, 1970, 3(3): 114.

[36] FITTS P M. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement[J]. Journal of Experimental Psychology General, 1992, 121(3): 262-269.

[37] MACKENZIE I, SCOTT. Fitts' Law as a Research and Design Tool in Human-Computer Interaction[J]. Human Computer Interaction, 1992, 7(1): 91.

[38] WINGRAVE C A, BOWMAN D A. Baseline Factors for Raycasting Selection[J]. Proceedings of Virtual Reality International, 2005(1): 1-10.

[39] POUPYREV I, WEGHORST S, BILLINGHURST M, et al. A Framework and Testbed for Studying Manipulation Techniques[J]. Proceedings of the ACM symposium on Virtual reality software and technology, 2013(1): 1379-1389.

[40] ARGELAGUET F, ANDUJAR C. Efficient 3D Pointing Selection in Cluttered Virtual Environments[J]. IEEE Computer Graphics and Applications, 2009, 29(6): 34-43.

[41] KOPPER R, BOWMAN D A, Silva M G, et al. A Human Motor Behavior Model for Distal Pointing Tasks[J]. International Journal of Human-Computer Studies, 2010, 68(10): 603-615.

[42] Teather R J, Stuerzlinger W. Pointing at 3D Targets in A Stereo Head-tracked Virtual Environment[C]. 2011 IEEE Symposium on 3D User Interfaces (3DUI), 2011.

[43] ALORAINI S M, GLAZEBROOK C M, Sibley K M, et al. Anticipatory Postural Adjustments During a Fitts' Task: Comparing Young Versus Older Adults and the Effects of Different Foci of Attention[J]. Human Movement Science, 2019, 64: 366-377.

[44] MEYER D E, ABRAMS R A, KORNBLUM S, et al. Optimality in Human Motor Performance: Ideal Control of Rapid Aimed Movements[J]. Psychological Review, 1988, 95(3): 340-370.

[45] STUART K, JOCK D, GEORGE G. A Morphological Analysis of the Design Space of Input Devices[J]. ACM Transactions on Information Systems (TOIS), 1991, 9(2): 99-122.

[46] Kytö M, Ens B, Piumsomboon T, et al. Pinpointing: Precise Head-and Eye-based Target Selection for Augmented Reality[C]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.

[47] 邓成龙. 虚拟现实中远距离放置任务的人类操作特性与模型[D]. 上海: 华东师范大学, 2019.
DENG Cheng-long. Human Operation Characteristics and Model of Long-distance Placement Task in Virtual Reality[D]. Shanghai: East China Normal University, 2019.

[48] GALEN G, JONG W. Fitts' Law as the Outcome of a Dynamic Noise Filtering Model of Motor Control[J]. Human Movement Science, 1995, 14(4): 539-571.

[49] MACKENZIE C L, MARTENIUK R G, DUGAS C, et al. Three-dimensional Movement Trajectories in Fitts' Task: Implications for Control[J]. Quarterly Journal of Experimental Psychology A, 1987, 39(4): 629-647.

[50] SHIMOJO S, SIMION C, SHIMOJO E, et al. Gaze Bias Both Reflects and Influences Preference[J]. Nature Neuroscience, 2003, 6(12): 1317-1322.

[51] RANGEl A, CAMERER C, MONTAGUE P R. A Framework for Studying the Neurobiology of Value-based Decision Making[J]. Nature Reviews Neuroscience, 2008, 9(7): 545-556.

[52] YANG S C H, LWNFYEL M, WOLOERT D M. Active Sensing in the Categorization of Visual Patterns[J]. Elife, 2016(5): 12215.

[53] AKINYELU A A, BLIGNAUT P. Convolutional Neural Network-Based Methods for Eye Gaze Estimation: A Survey[J]. IEEE Access, 2020, 8: 142581-142605.

[54] CUBERO G, CARLOS, REHM M. Intention Recognition in Human Robot Interaction Based on Eye Tracking[C]. Springer, Cham: FIP Conference on Human-Computer Interaction, 2021.

[55] KRAJBICH I, ARMEL C, RANGEL A. Visual Fixations and the Computation and Comparison of Value in Simple Choice[J]. Nature Neuroscience, 2010, 13: 1292-1298.

[56] SMITH S M, KRAJBICH I. Gaze Amplifies Value in Decision Making[J]. Psychological Science, 2019, 30(1): 116-128.

[57] GLUTH S, NKORTMAN N, MVITALI C. Value-based Attention but Not Divisive Normalization Influences Decisions with Multiple Alternatives[J]. Nature Human Behaviour, 2020, 4: 634-645.

[58] JANG A I, SHARMA R, DRUGOWITSCH J. Optimal Policy for Attention-modulated Decisions Explains Human Fixation Behavior[J]. eLife Sciences, 2021(10): 1-19.

[59] KOOCHAKI, FATEMEH, LALEH N. Predicting Intention through Eye Gaze Patterns[C]. IEEE Biomedical Circuits and Systems Conference (BioCAS), 2018.

[60] KOOCHAKI F, NAJAFIZADEH L. A Data-Driven Framework for Intention Prediction via Eye Movement With Applications to Assistive Systems[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2021, 99: 1-1.

[61] OVIATT S. Advances in Robust Multimodal Interface

Design[J]. IEEE Computer Graphics and Applications, 2003, 23(5): 62 - 68.

[62] PFEUFFER K, ALEXANDER J, MING K C, et al. Gaze-shifting: Direct-indirect input with Pen and Touch Modulated by Gaze[C]. Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, 2015.

[63] SHUMIN Z, CARLOS M, STEVEN I. Manual and Gaze input Cascaded (MAGIC) Pointing[C]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1999.

[64] YANG Bo. Sliding Gaze: Head Movement and Touch Gesture Based Target Selection Technology[D]. Tokyo: Waseda Universit, 2019.

[65] JANI K, MASTANEH TA, AHMED H, et al. Automated Tool Detection with Deep Learning for Monitoring Kinematics and Eye-hand Coordination in Microsurgery[J]. Computers in Biology and Medicine, 2022, 141(188): 105121.

[66] CORDEIRO D, CORREIA N, RUI J. ARZombie: a Mobile Augmented Reality Game with Multimodal Interaction[C]. 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN), 2015.

[67] MARDANMEGI D, MAYER B, PFEUFFER K, et al. Eyeseethrough: Unifying Tool Selection and Application in Virtual Environments[C]. IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2019.

[68] BAI H, SASIKUMAR P, YANG J, et al. A User Study on Mixed Reality Remote Collaboration with Eye Gaze and Hand Gesture Sharing[C]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

[69] LI Y, CAO Z, WANG J. Gazture: Design and Implementation of a Gaze Based Gesture Control System on Tablets[J]. Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies, 2017, 1(3): 1-17.

[70] SOUSA M, DOS ANJOS R K, MENDES D, et al. Warping Deixis: Distorting Gestures to Enhance Collaboration[J]. Conference on Human Factors in Computing Systems - Proceedings, 2019(1): 1-12.

（上接第 26 页）

[96] 薛澄岐. 人机融合、智能人机交互、自然人机交互未来人机交互技术的三大发展方向——薛澄岐谈设计与科技[J]. 设计, 2020, 33(8): 52-57.
XUE Cheng-qi. Human-machine Interation, Intelligent Human Computer Interaction, Natural Human Computer Interaction Three Development Directions of Human Computer Interaction Technology in the Future: Xue Chengqi On Design And Technology[J]. Design, 2020, 33(8): 52-57.

[97] 张青. 工业设计中多重感官交互增强现实系统设计与应用[J]. 科学技术与工程, 2018, 18(32): 206-211.
ZHANG Qing. Design and Application of Multi Sensory Interaction Augmented Reality System in Industrial Design[J]. Science Technology and Engineering, 2018, 18(32): 206-211.

[98] 陈建华, 崔东华, 罗荣, 等. 军事指控系统多通道人机交互技术[J]. 指挥控制与仿真, 2019, 41(4): 110-113.
CHEN Jian-hua, CUI Dong-hua, LUO Rong, et al. Multi-modal Interaction Technology of Military Command and Control System[J]. Command Control & Simulation, 2019, 41(4): 110-113.

[99] ZAHRA EMAMI, TOM CHAU. The Effects of Visual Distractors on Cognitive Load in a Motor Imagery Brain-Computer Interface[J]. Behavioural Brain Research, 2020, 378: 10.

[100] SON JOONWOO, PARK MYOUNGOUK. The Effects of Distraction Type and Difficulty on Older Drivers' Performance and Behaviour: Visual vs. Cognitive[J]. International Journal of Automotive Technology, 2021, 22(1): 10.

[101] L CAI, J DONG and M. WEI. Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning[C]. 2020 Chinese Automation Congress (CAC), 2020(1): 5726-5729.

[102] ZENG Bo-tao, FENG Zhi-quan, XU Tao, et al. Research on Intelligent Experimental Equipment and Key Algorithms Based on Multimodal Fusion Perception[J]. IEEE ACCESS, 2020(8): 10.