元宇宙需要人机交互的突破

文 / 史元春

摘　要：元宇宙目标实现万物的信息化和智能化，创造一个信息充分包围人的虚实融合空间，演化生成时空无界的新型社会形态。人机交互是元宇宙的核心关键技术，人机接口的扩展和虚拟化，实现人机之间高效交换语义信息技术挑战大。掌握人机交互科技优势，对推动相关产业发展有着至关重要的作用。本文分析元宇宙人机交互的挑战，重点探讨交互意图推理的突破思路与最新进展。

1 元宇宙意味着什么

刚刚过去的 2021 年被称为元宇宙（Metaverse）元年，主要是由一系列商业事件引起的，上半年游戏公司罗布乐思（Roblox）上市股价大涨，继而在中国重奖推动元宇宙创作罗布乐思全国创作大赛（Roblox National Awards 2021）；社交媒体公司脸书（Facebook）在多年前收购和重塑了虚拟现实产品 Oculus 后，于 10 月直接更名为 Meta，彰显其打造元宇宙新型计算平台的决心；微软也携其打造多年的混合现实产品 Hololens，发布了数字化生产力平台 Mesh for Microsoft Teams。元宇宙突现在 2021，与大众在近两年的疫情中越来越依赖数字化平台不无关系，更期待沉浸感、交互性、打破物理约束、进入平行孪生环境的体验升级，也反映了不断进步的信息技术融合可能带来的超预期变革。

Metaverse 一词来自于美国著名科幻作家 Neal Stephenson 1992 年发表的小说《雪崩》（Snow Crash），被描述为：“戴上耳机和目镜，找到连接终端，就能够以虚拟分身（Avatar）的方式进入由计算机模拟、与真实世界平行的虚拟空间。”30 年前产生如此充满想象力的科幻作品，与当时的科技发展有很大关系。

《雪崩》中的元宇宙是虚拟现实的，虽然 VR（Virtual Reality）一词在 20 世纪 80 年代才出现，而虚拟现实之父 Ivan Sutherland 早在 1965 年就发表论文 The Ultimate Display，1968 年还研制成功了带跟踪器的头盔式立体显示器（Head Mounted Display，HMD）。三维图形生成、多传感器交互和高分辨率显示等 VR 技术在 80 年代蓬勃发展，1987 年已经出现在大众科学杂志《科学美国人》的封面。基于 VR 科技基础，曾经是程序员的作家 Neal Stephenson 对未来的构建充满想象力。

随着 80 年代后期 PC 的出现和普及，关于计算机的未来思潮与实践，不仅有 VR 的虚拟化，还有“普适计算”的现实化，这就是以 1991 年也是发表在《科学美国人》的施乐研究院 PARC 的计算机首席科学家 Mark Weiser 的“The Computer for 21st Century”提出的 Ubiquitous Computing。他指出，计算将从桌面 PC 扩展到移动终端，融入到物理空间泛在设备上，动态组成面向用户连续活动的易用系统，并预言他文中 PARC 研制的典型个人移动终端 Tab 将在 20 年后成为主流后，互联起更多的不为用户关注的设备（物联、穿戴等嵌入式设备），实现

Embodied Virtuality（EV）——融入现实。诚如所言，十多年后以智能手机为终端的移动计算普及到日常应用。Weiser 当时还明确指出，他提出的 EV，是与那时很热的 VR 完全不同的理念（见图 1）[1]，VR 是将现实虚拟化到计算机中，而 EV 或者说普适计算是将信息和计算融入到现实世界中去。

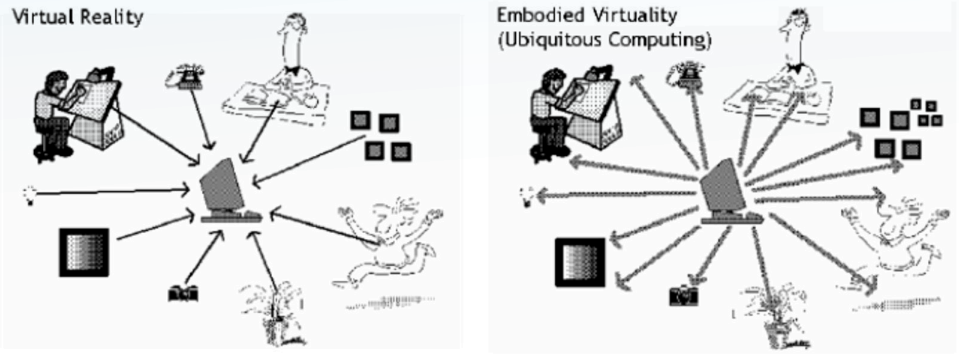


图 1　Mark Weiser 阐释 VR 与 EV 是方向相反的计算技术发展路线

之后 30 年，VR 和 UC 代表的终端技术都蓬勃发展成为可供用户选择的产品，虚拟化和现实化的应用交融成为虚实融合的新的热词——元宇宙。这个新热的词不同于之前曾经大热的技术术语（如云计算、物联网、人工智能等），甚至很难说是一个技术术语，但它的背后是扎扎实实数十年的信息技术进步，也彰显着信息技术的巨大力量。宇宙在 2 000 多年前的《淮南子·齐俗训》就有了明确定义："四方上下谓之宇，古往今来谓之宙"；而 Mata（元）是关于时空无界的万事万物的本源，唯有信息科技方可描述之、计算之、变换之、作用之。

信息科学在上世纪三四十年代奠基，十年左右的时间出现了图灵机、冯·诺伊曼结构、信息论、控制论；数十年的时间，信息技术的终端——计算机，从躺在实验室、到立在办公桌、到握在掌上，逐渐成为人们日常生活的一部分。信息科技之重要，是因为它撬动了自然界的信息流，自然界的物质、能量和信息三大要素，人们首先认识了物质（材料），然后认识了能量（动力），最后才认识了信息（和知识）；认识了信息，拥有了信息处理、传播、生成的工具，人们不仅能够更好地认识自然界，还能够更好地认识人类社会和主观世界，并架起主客观世界的桥梁。元宇宙因其虚实融合的特性而能以前所未有的灵活方式，对真实社会的经济和思想产生影响。

2 元宇宙需要怎样的人机交互技术

元宇宙的目标是实现万物的信息化和智能化，创造一个信息充分包围人的虚实融合空间，演化生成时空无界的新型社会形态。元宇宙是信息技术蓬勃发展的集大成应用环境，我们可以从图 2 所示[2]的这张著名的元宇宙价值链图上看到诸多当今先进的 IT 技术，如 5G/6G、穿戴、眼镜 XR、AI、区块链、3D 引擎等。元宇宙是面向终端用户（2C）的，所以我们看到，人机交互（HCI）是核心关键技术。

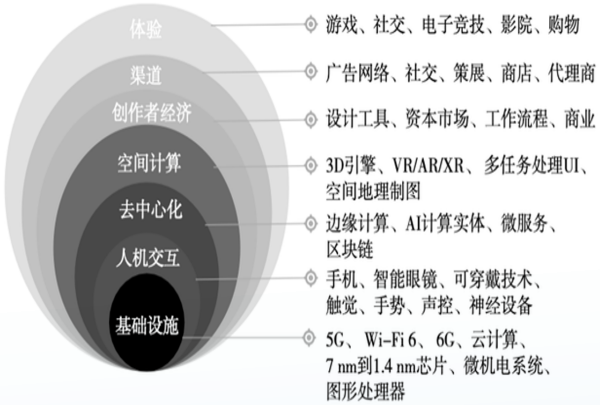


图 2　Metaverse 的七层价值链

1 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.51&rep=rep1&type=pdf
2 https://www.163.com/dy/article/GJ9LMF4005310573.html

人机交互研究人机之间高可用的信息交换原理和技术，建立起计算机多种模态的输入输出软硬件交互接口，及其构成的用户终端界面，形成特定的交互模式。计算机发展历史上，交互模式从 CUI（字符用户界面）的键入命令，到 GUI（图形用户界面）的鼠标拖拽，再到智能手机上的动作语音操控，不断突破用户使用计算机的难度瓶颈，计算得以从少量专家用户的科学计算扩增到数十亿人的日常应用。人机交互也是元宇宙系统的基础能力，其性能水平直接决定了人在元宇宙中的能力边界，进而决定了元宇宙对人类的价值。

元宇宙目前有成熟可用的人机交互技术吗？没有。首先，元宇宙的用户终端将从智能手机扩展开来，新型的终端将多种多样，典型的包括智能眼镜和智慧物联网环境，如家居等，这些新型终端在交互接口上的共性是接口的虚拟化、远程化和多映射关系，不同于 PC 和手机的固定接口并能提供明确的反馈确认，目前实现基本的交互功能尚有困难。

交互技术在不断进步，输入技术上，则都需要实现交互接口上的操控目标对象（对象空间位置明确，又称为可视目标）、表达抽象命令（不是基于目标位置的），以及输入语言内容三个原子功能，表 1 简要归纳了历史上几种主要的交互模式是如何实现这些原子交互功能的。

表 1 输入接口原子功能的实现

输入接口原子功能	接口模态	CUI 字符用户界面	GUI 图形用户界面（PC）	GUI 图形用户界面（手机）
操控目标对象	视觉搜索、输入设备 / 人体动作访问	无	鼠标点击	手指触摸
表达抽象命令	语言、行为编码	命令行（命令的语言编码）	菜单 / 快捷键	触屏手势 / 语音命令
输入语言内容	键盘、手写、语音	物理键盘	物理键盘 / 手写板	软键盘 / 语音 / 手写

简要讨论下智能眼镜。眼镜的意义是解放双手，沉浸三维，回归自然，但自然交互上，由于失去了精准的输入设备（如桌面的鼠标），缺少触觉反馈的感知支持（如手机触屏），完成这三类基本的交互功能，一直是智能眼镜固有的难题：① 视觉注意过度造成速度慢和眩晕。3D 对象的访问一般都需要人眼视觉参与，但我们用手访问精细的 3D 对象时，视觉注意力的高度集中和随动图像难以准确匹配人的视觉感知，不仅访问速度慢，还是造成用户眩晕的重要因素。② 手势动作的设计空间虽大但可用性不高。手势到命令的映射关系设计和用户表达需要符合人的认知和表达能力，同时还存在传感和识别技术上的困难。③ 文本输入必要但困难。用手势在空中访问虚拟键盘输入文字时，速度慢到不及手机的 1/3；而语音并不万能，出于隐私考虑，很多场合不便出声；并且缺少连续交互能力，每次都需要唤醒。因此，这些问题限制了用户的表达能力，是智能眼镜一直难以成为通用用户终端的症结所在；为解决这些问题，市场上还出现了专属折叠蓝牙键盘和鼠标这样不得不“回归”的配置，完全失去了眼镜解放双手的意义。

进一步地，元宇宙空间，人机关系将发生重大变革，机器从被动应答者向主动服务者身份转变，交互从单一显式的用户动作表达向隐式机器智能推理与显式用户表达融合方向发展，在恰当时间与情境下提供用户亟需的智能服务成为必然发展趋势。也就是说，人机交互的路径，将从现在用户记忆搜索应用和界面的模式，转换到机器主动感知和推送服务，极大缩减交互路径的模式。由于用户终端多样，这一层面的问题还反映在同样功能的服务需要设计实现为每种终端上的应用功能的开发难题。

下面主要简述在交互接口难题上的研究进展。

3 自然动作交互意图识别

自然人机交互（NUI）是人机交互领域早就被提及，一直在追求的目标。在 GUI 发端的上世纪 60 年代，图形化、可视界面搜索代替主要依靠记忆的命令行方式，由于记忆和表达负担的降低，就被称为是自然的；普适计算被提出的 90 年代初，预测信息技术将附在很多场景服务人的日常应用，自然交互的研究不仅要更多地降低记忆和表达负担，在接口上还体现为多种“解放式”（free）的特征：device-free 指脱离专门的输入设备（如鼠标、笔）而直接用人体器官作为输入工具，触屏就是以手指为输入工具；eyes-free 指无需或减少交互中的视觉注意力；hands-free 是解放双手，包括手上不持握设备和不用手交互。在现实空间和虚拟空间中连续、自然的动作交互，将是元宇宙的主流交互方式，但难题是，如何在模糊的自然行为数据上推理人的交互意图?

3.1 动作交互的意图推理问题

动作数据包括手指、手部、头动及身体运动等，是当前用户表达交互意图的主要通道。传感动作的传感器多样，针对动作的特定观测点按照一定的时间间隔采集数据。对于每次采样，每种传感器会获得一个拥有 M 个分量的数据，如果有 N 个独立的传感器，交互接口对其就形成 $M\times N$ 个观测点。同时将时间范围内所表达的交互意图记为 Y，动作交互意图的识别过程可以形式化地表达为

$$\begin{bmatrix} S_1^1 & S_2^1 & \cdots & S_N^1 \\ S_1^2 & S_2^2 & \cdots & S_N^2 \\ \vdots & \vdots & & \vdots \\ S_1^M & S_2^M & \cdots & S_N^M \end{bmatrix} \rightarrow Y \tag{1}$$

这是一个如何根据观测到的用户交互信号 S 来推断用户交互的意图的推理问题。问题的挑战主要在于难以克服的不确定性：① 在完成一项动作时，用户对自身身体运动的规划往往是确定的，但每次做出的动作又不完全相同，体现在动作完成过程中幅度范围、时长的偏差。如图 3 所示[1]；② 不同用户针对同一交互意图可能产生不同的交互行为，如在 QWERTY（软）键盘输入文本时，用户指法可能不同。

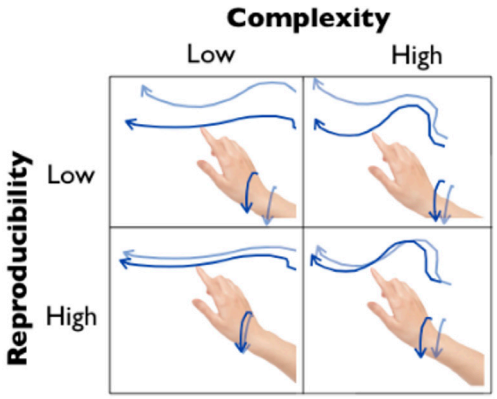


图 3 人体动作的信息容量：重复性和复杂度

3.2 意图推理中的智能算法

由于观测用户行为的方式不同（传感器、观测位置），以及针对的应用场景不同，意图推理问题的输入和输出也不同，一般可以将意图推理问题分为分类和回归两类问题，前者需要根据用户输入的信号，从多个可能的类别中找到对应的分类（如状态检测）；后者需要根据交互信号来计算某一具体的数值或指标，以达到提升精度等目的。可采取的方法包括下面四种[2]。

（1）模板匹配类算法是最简单的分类方法，核心是模板匹配中的相似度的计算，松弛的计算方法通过动态规划在所有可能的匹配中找到相似度最高的一个[3]。此方法在分类问题的类别较少，混淆性相对较小的情况下，具有非常好的效果。而且其计算速度很快，因而常常被用于滑行文本输入、动作和手势识别、视线焦点的位置计算等任务。但若模板

1 https://www.docin.com/p-1611180733.html
2 https://www.doc88.com/p-3846108342332.html?r=1
3 https://eprints.lancs.ac.uk/id/eprint/84705/1/EaglseSenseCHI2017.pdf

数量过多，或者模板之间的相似性太大，其准确性和效率都会受到明显影响。

（2）决策树也是一类非常简单而有效的模型，具有直接的物理意义和高效的运行速度。其缺点是在数据量不足时容易出现分类不准；在数据充足时，又容易出现过度拟合。而随机森林的方法则提升了决策面的维度[1]。两类方法常被用于视线焦点位置推测、动作识别、操作手指区分、区分滑动操作是否正确、学习语言的规则等任务中。

上述“白盒子”算法来源于人们对客观规律的数学建模，具有较强的推广性。难度在于建模和适用性。

（3）隐变量机器学习类方法往往利用包含不可见状态的随机过程或神经网络来对输入进行建模，支持状态机 (SVM)、隐 Markov 模型 (HMM) 都被广泛用于各种分类问题中 (如手势识别、身份区分、握持姿势识别)[2]，具有计算速度快、准确性高的优点；隐变量机器学习类方法利用多层的“神经元”实现描述性极强的模型，在基于图像的分类和回归等问题上具有极高的准确性。

这些“黑盒子”算法不需要大量的人力建模，但强烈地依赖于训练数据。

（4）贝叶斯算法是基于统计学的一种分类算法。它基于概率论的知识，利用从观测数据中统计出来的特征量构建模型。模型参数一般具有直接的物理意义。在实践中，贝叶斯方法常常能实现高效的分类效率，同时拥有不输于神经网络等方法的分类准确性。相比于其他方法，贝叶斯方法针对较小的样本能产生更加准确的结果，而且能在结果之外同时计算出结果的置信度和显著性。

这是“灰盒子”算法，结合了黑白盒子的特点，通过概率统计方法将人的知识引入到算法模型中，对于无法确定的变量、关系，通过黑盒子的方法来完成。其同时具有对规律的可解释性和对数据的忠诚性。

3.3 基于贝叶斯的交互意图推理框架

我们提出了基于贝叶斯推理的交互意图推理框架（见图 4），将任务 - 情境模型 $P(I|CT)$ 和行为编码模型 $P(G|I,CT)$ 作为先验知识代入贝叶斯推理，可以走出单纯依靠感知数据 $P(G'|G)$ 进行交互语义识别所难以克服的数据不充分和数据 - 意图难对应的困境，大幅提高意图识别 $P(I|G',CT)$ 的准确性。基于该方法我们为智能手机研制了握持意图识别、软键盘容错输入等国际领先的产品技术，前者有效解决全面屏高误触难题；后者显著提升了预测纠错能力、提高了输入速度，支持着 7 亿多用户的日常应用。

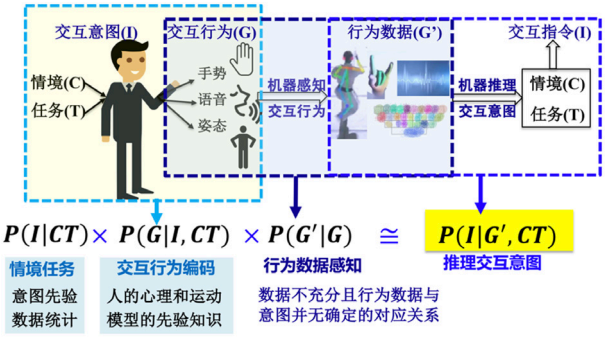


图 4 基于贝叶斯的交互意图推理框架

如图 4 所示，该推理框架在不同的动作、任务、情境下，几个模型有其对应物理意义的数据来源。下面以空中打字（Air Typing）为例，简介一下具体实现，这是元宇宙中很科幻的一项技术，也是第二节中语言输入功能的一个有效的解决方案。

（1）空中打字

空中打字问题的输入是用户产生的一系列落点 I，以及产生这些落点时，每个手指的点击幅度 D。我们需要计算的 $P(W|I,D)$ 就是图 4 中的交互意图，即在给定 I 和 D 的情况下，输入的目标单词是 W 的概率，其中 W 是任意一个可能的单词。该条件概率正比于 $P(W)P(I|c)P(D|c)$，其中 $P(W)$ 是语言模型，一般使用其词频来计算，是由当前的任务情境决定的；

1 https://dl.acm.org/doi/10.1145/2851581.2889430
2 https://dl.acm.org/doi/epdf/10.1145/2807442.2807504

$P(I|c)$ 衡量了用户的目标按键和落点实际位置之间的概率关系，可以用三维空间触摸模型来计算；$P(D|c)$ 衡量了用户的目标按键和使用的主动手指之间的概率关系，可以用空中多指联动模型来计算 (见图 5)。

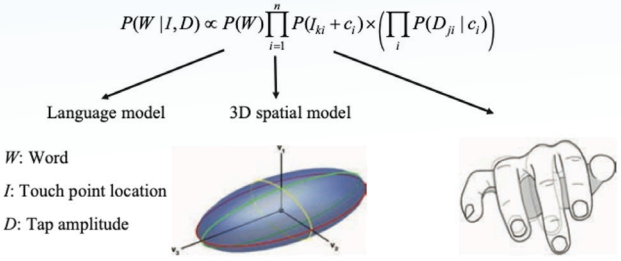


图 5 空中打字识别

通过结合语言模型、触摸模型和空中多指联动模型，贝叶斯推理框架同时计算了落点的位置信息、手指的指法信息和语言本身的信息，从而大幅度提升了输入准确率。用户实验表明，优化的算法首次实现了空中双手盲打的输入体验，可以使得用户在接近 30 WPM（每分钟 30 个单词）的输入速度（手机输入的平均速度）下，达到近 100% 的输入准确率，在空中打字这个已存在数十年的交互概念上实现了技术突破 [1]。

下面简介在此框架下元宇宙空间另两个原子交互任务的解决方案。

（2）可视对象无注视访问

如前文所述，虚拟空间中 3D 对象的访问是一个基本的交互功能，但视觉注意过度是访问速度慢和产生眩晕的一个主要原因，我们研究了无视觉参与完成虚拟目标抓取的技术 [2]。用户行为研究显示，用户依赖空间记忆和自体感知能力，可以在不进行视觉搜索的情况下，抓取目标位置。显然，在这种情况下用户的抓取位置是存在系统偏差和随机误差的。我们量化了用户抓取动作的系统偏差和随机偏差的概率关系，在解码时进行了补偿（见图 6）。通过线性差值来模拟在整个空间中的偏差分布。在此基础上首创了无需视觉注意的空中目标选取技术，虚拟目标抓取速度较 Hololens 的有视觉反馈的 1.325 s 降低至 0.988 s，并有效降低眩晕感。

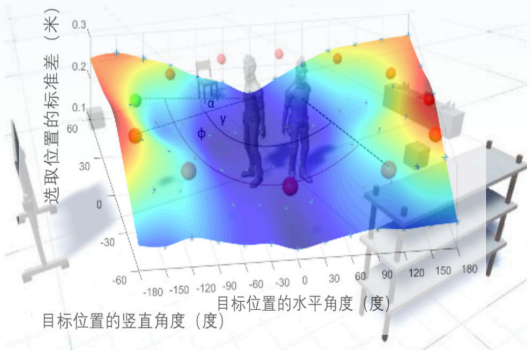


图 6 用户动作控制偏差补偿和布局优化

（3）基于交互语义的动作编解码

用动作（主要是手势）表达命令有着 device-free 和 eyes-free 的极大优势，但自深度视觉传感技术成熟、手势识别具有可行性以来，已经有十多年的时间，手势的使用极其有限，主要原因有两个，一是人的连续动作数据中，有交互意图的有意动作与无意动作都“内嵌”其中，提取困难（见图 7），而根据图 4 的推理框架，复杂如空中打字这样的连续动作，只要能建立起基于自然动作的内在时空特征模型就可找到可行解；另一个原因则是动作编码本身，需要有很低的发现成本、完成成本和记忆成本，需要深入研究人的认知特点，核心问题是根据场景任务建立“动作 - 指令”的映射关系。例如，元宇宙中访问菜单很低效，在游戏装备和办公工具的选择任务上，我们设计了基于已有经验的动作手势，可直接表达和获得相应的对象，如需要弓箭就做拉弓动作、需要笔就做写字的动作 [3]。语音交互的自然唤醒，凑近和捂嘴行为本身携带语音交互意图，并可计算出行为的声音等多模态信号特征；保证可触发和高准确性的同时，凑近还允许更小的声音，兼顾了隐私性，正在转化为手机和穿戴设备上鲁棒的语音交互免唤醒词功能。

1 https://dl.acm.org/doi/10.1145/3173574.3173616

2 https://www.sciencedirect.com/science/article/pii/S2096579619300233

3 https://pi.cs.tsinghua.edu.cn/publication/page/2/

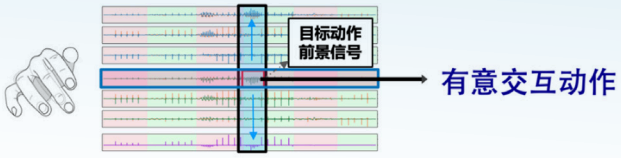


图 7 在连续动作数据中识别有意交互动作

事实上，虽然手势应用最广，但是在一些情况下我们不能使用手部进行交互，比如由于身体、任务和情景的影响，手势模态受损、受限和不适，难以进行交互，身体的其他部位也能表达交互语义。图 8 所示是我们在图 4 框架下实现的系列动作交互技术[1]。

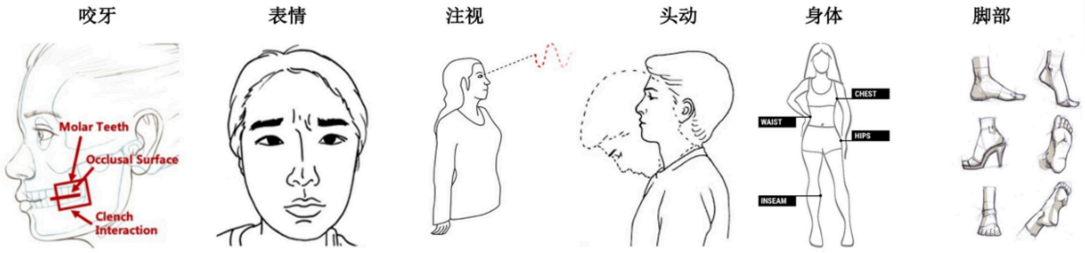
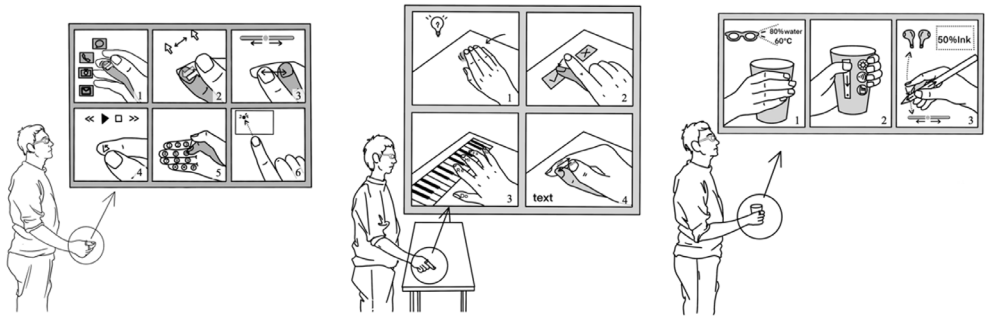


图 8 手势之外的交互动作

（4）戒指交互：微手势精准小型穿戴交互设备

前述三类交互技术分别针对了一两个原子交互功能的实现，且都不需要用户操控任何设备，用户实验也都显示了其设计功能上很高的可用性和效率，而研制精准的操控设备也是虚拟化空间交互的一种有效的解决方案。例如，VR 智能眼镜上的手柄可准确访问可视对象，但抽象命令表达和文本输入功能不足。我们最新研制了精准的小型穿戴交互设备 DualRing[2]，利用人最灵活和准确的输入器官——手指，戴在拇指和中指上可感知手指运动关系的戒指，不仅可以感知相对于表面的绝对手势，还可以感知手节之间的相对姿势和运动，如图 9 所示的自然手势交互的设计空间。用户实验表明，DualRing 可以实现 2D、3D 对象准确访问、手势操作，甚至打字，其可用性、效率和新颖性受到用户的青睐，尤其是在综合手势方面具有显著优势。



(a) 手内交互　(b) 手与面交互　(c) 手与物交互

图 9 智能戒指的手势设计空间

4 结束语

元宇宙中用户交互行为无处不在且多模连续，导致交互数据模糊，为交互意图准确推理带来巨大挑战。无论是扩展现实的智能眼镜还是智慧互联的现实环境，交互接口都不再是单一固定的界面设备，在虚拟化的空间上如何实现操控目标对象、表达抽象命令、输入语言内容三个输入交互接口上的原子功能，目前尚无高可用的产品技术，是人机交互领域的热点研究方向，核心是动作交互的意图推理问题，可采用的智能计算方法涉及白盒、黑盒和灰盒。

1https://www.sciencedirect.com/science/article/pii/S2096579619300233

2 https://dl.acm.org/doi/10.1145/3478114

本文重点介绍了我们提出的基于贝叶斯的自然交互意图理解计算框架，具有若干优点：首先，其构造具有可解释性，对其结构和子项的进一步研究是原理性的，有助于产生科学发现和规律解释；其次，其求解方式是知识结合数据的，很多情况下，只需要相对较少的数据就可以获得满意的解，这非常符合人机交互新技术研发的需求，因为在技术部署之前是难以收集大量数据的。最后，其结果具有推广性，或者说所获得知识经验是可以迁移的。

交互路径的固化模式，是元宇宙人机交互面临更高层面的难题，机器需要从被动应答者向主动服务者身份转变，对用户特征（如残障或操控条件导致的信道不可用）和用户任务的适应，需要建立在分布式多模态情境感知推理和交互任务知识图谱的基础上，限于篇幅，本文未展开论述。

（参考文献略）

史元春

清华大学人工智能研究院智能人机交互中心主任、普适计算教育部重点实验室主任、计算机系“长江学者”特聘教授。主要研究方向为人机交互、普适计算、多媒体等。曾获得两项国家科技进步奖二等奖。CAAI 智能交互专委会副主任，CAAI/CCF Fellow。