第5章 多模态输入技术

5.1 多模态输入概述

为了和用户进行高效互动，增强现实系统的输入与人的多模态感官系统紧密耦合。目前增强现实主流的专业设备有微软的 HoloLens、Magic Leap One、Google Glass 等。这些设备普遍可以支持手势、语音交互操作。在手持终端的手机平板上，摄像机、麦克风、触摸屏成为基本输入，而体感（包括眼神）信号也日渐成为包括专业设备和手持终端上的标配功能。事实上，单一模态的输入无法满足复杂多变的真实场景需求。根据场景和任务，允许用户选择合适的模态，甚至智能组合多种模态是高效、准确地理解真实世界和用户意图的不二选择。

5.2 键 盘 输 入

打字是一项与人体工程学息息相关的任务，是现代人生活中不可或缺的一部分。人们通常花费大量时间在键盘打字上，对键盘输入方式的研究也日益增多。键盘的位置是影响用户打字舒适度的重要因素。增强现实逐步走入人们的生活或者工作场所。在这样的情况下，有若干解决方案可以允许用户继续使用键盘作为输入工具。

- 利用一些外设（如智能手表、小型的蓝牙键盘等）替代键盘。这并不是最佳方式，因为对外设的依赖会降低增强现实应用的独立性。
- 基于手持终端（手机和平台）的应用一般可以调用屏幕上的虚拟键盘。这是对开发者最简单，对用户也是最容易接受的方式。
- 基于头戴式增强现实硬件的应用，若要实现全键盘的输入方式，则需要探索新的技术。最理想的方式是，系统能将键盘投影到用户视野中，并通过体感交互的方式允许用户快捷地输入文字。

针对最后一个情况，实际上有很大难度和不少尚未解决的技术难点。在头戴式显示器中，来自真实世界的布局干扰、颜色混合和投影位置等问题都将影响用户使用该虚拟键盘的适用性。深度顺序、对象分割和场景失真都容易导致用户难以通过透明的头戴显示器查看内容。这些问题会影响虚拟键盘按键的可读性和可见性。因此，用户需要改变原有的平面输入方式，开发者需要发明更好的输入方式来适应增强现实的硬件平台，能够方便用户输入文字。

增强现实输入方法相比原始的方法确实存在一些问题，具体区别如图 5.1 所示。其中，

图 5.1(a)所示为手和显示器在标准的位置进行文本输入；图 5.1(b)所示为使用 AR 设备时，用控制器进行文本输入；图 5.1(c)所示为使用 AR 设备时，使用手部光学追踪进行文本输入。

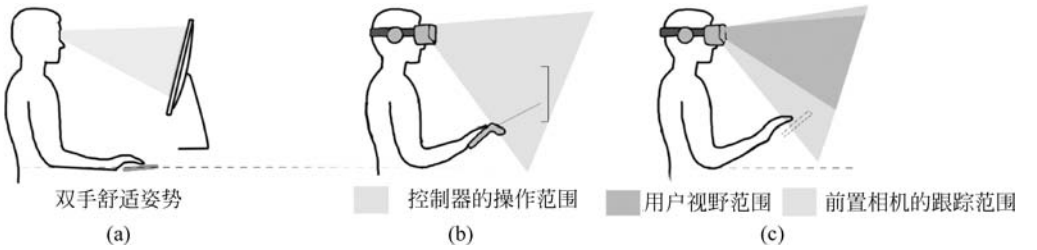


图 5.1　增强现实输入与普通键盘、手控传感器的比较

在图 5.1(a)中，用户的肘部是水平的，这个姿势使用户可以放松手部。而对于图 5.1(b)和图 5.1(c)，用户为了符合设备的输入方式，不得不将手抬起。长时间保持这样的姿势必然会使用户疲劳，尤其手握 AR 设备的控制器相比空手更加劳累。

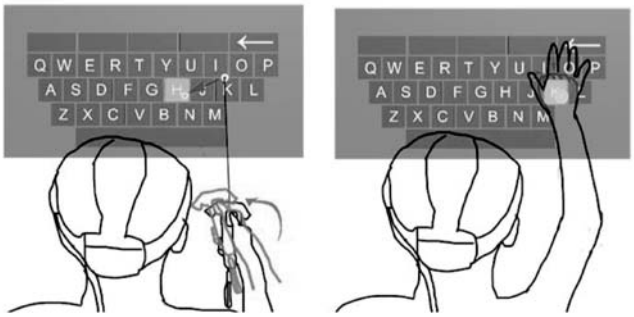
不仅是键盘的位置会给用户带来影响，键盘的形状和尺寸也会带来重要的影响。弯曲的键盘更加符合人体工程学(见图 5.2(a))，而人们已经习惯了普通的键盘(见图 5.2(b))。键盘的大小也应当合适，太小则用户无法看清按键，而太大则用户无法看清键盘全貌，不得不转动头部去看键盘的各个部分。

图 5.2　不同键盘的形状

基于手持终端屏幕呈现的虚拟键盘是增强现实应用的标准输入方式。和以 Google Glass 为代表的专业设备相比，在屏幕上的虚拟键盘是主流用户都能够接受、快速上手的方式。相反，Google Glass 等专业设备则需要通过眼神、手势、控制器等方式将指针移动到目标键位上，再确认进行输入该字符，这种方式是非常缓慢的，需要花费大量的时间。与此同时，将键盘保持在视角中并且长时间利用手势、控制器容易导致疲劳，大幅降低了用户的体验。

与虚拟环境及其对象进行交互的最常见方式之一是通过手持控制器(见图 5.3(a))。该设备使用从其投射到虚拟环境的射线作为指向机制。射线的末端类似于光标。用户只需移动控制器以指向所需字母即可在虚拟键盘上打字。选择可以通过按键输入或滑动输入完成。另外一种方法是让用户仅用手与虚拟键盘进行交互(见图 5.3(b))。通过头戴式设备的前置摄像头捕获手掌和手势的位置。也就是说，用户使用手掌在空中的位置来指示光标的位置，触发虚拟键盘的文字输入。用户根据他们的手在虚拟键盘上移动光标。

部分虚拟键盘的变种，可以有效降低原有的缺陷。例如 Punchkeyboard(见图 5.4)，这



(a) 手持控制器交互　　(b) 手掌位置控制光标交互

图 5.3　虚拟环境中的交互方式

个项目允许用户使用两个手柄控制器，像双手正常按键一样，使用控制器的虚拟指针快速在虚拟键盘上按键，提升用户的打字体验。这个项目同时通过自动的单词补全和下一个词语预测来加速文字输入的速度。

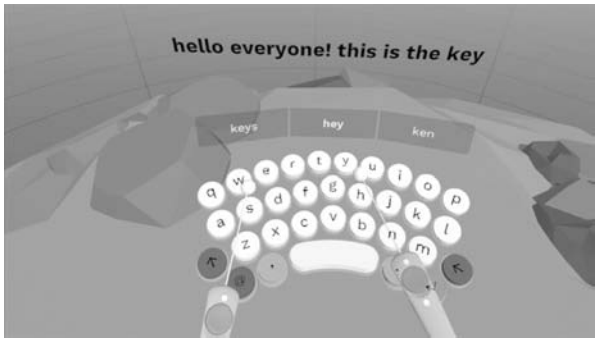


图 5.4　Punchkeyboard 项目

手部跟踪键盘(见图 5.5)和实体键盘的输入方式基本一致，唯一的区别是虚拟键盘代替了实体键盘，所以用户无须任何学习就可以直接使用。这种输入方式明显地提高了打字速度，能够和基本输入方式的速度相同。但也存在一些缺陷，手部位置需要在摄像头范围内，可能会导致如前文所述的疲劳姿势。如果纯空中手势无法提供物理触觉反馈，用户则会产生一种没有反馈的失落感。

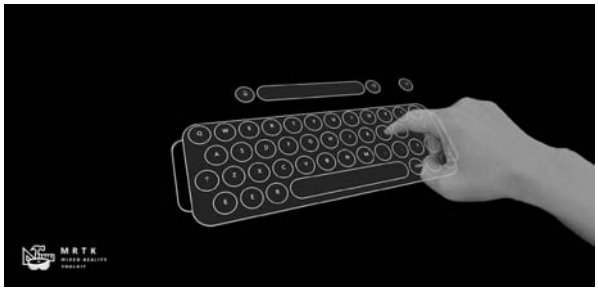


图 5.5　Microsoft Mixed Reality ToolKit 手部跟踪键盘

5.3 语音输入

语音输入是通过计算机自动地将人类的语音转换为相应的文字的技术。随着语音识别和自然语音处理技术的发展，使得在增强现实中能够使用语音进行交互。语音是适合增强现实的一种输入方式。特别是针对头戴式设备，没有触摸屏，一般只能通过眼镜的镜腿区域内有限数量的按键进行交互。语音交互可以摆脱这个局限性，更好地解放双手。

现在已经有成熟的技术让人们和电子设备进行语音沟通，如科大讯飞、腾讯、阿里巴巴、华为、苹果、微软等公司，都有成熟的语音识别软件提供给开发者。微软已经在增强现实设备 HoloLens 上成功应用语音识别，国产增强现实眼镜也普遍搭载了麦克风（见图 5.6），用于收集语音输入信号。

图 5.6　国产品牌 Rokid 增强现实眼镜的麦克风展示

人机语音交互有以下五个关键处理阶段。

（1）机器接收到用户语音后，首先通过语音识别将语音转换为文本，并且可保留语速、音量、停顿等语音本身的特征信息。

（2）机器通过自然语言理解从文本中理解用户意图。

（3）机器通过对话管理决策接下来的动作，并更新对话状态。

（4）机器通过自然语言生成将决策后的动作生成为回复给用户的文本。

（5）最后，机器通过语音合成将回复给用户的文本转换为语音，完成一次交互。

要启动语音输入，第一个步骤是要实现语音唤醒，即通过特定词语启动增强现实设备，使其进入理解人类自然语言的模式。这个看似“多余”的动作是为了允许设备在长时间不交互的情况下，能够进入休眠、低功耗状态。这个功能在主流提供语音交互的设备上均有提供。语音唤醒普遍是针对已经绑定的用户，非绑定的用户即使知晓该唤醒词也无法启动设备的语音交互功能。

语音输入能够方便普通人，特别是部分有运动障碍的用户使用。相比动手而言，动嘴相对轻松、自然。这是一种方便快捷地用于输入文本以及控制系统和应用程序的方法。对于文本输入，正常人说话的速度明显快于用手进行输入。对于系统控制，语音对于界面中多个按钮的选择非常快。因为原始的界面控制中，触发一个命令需要三个步骤：用眼睛寻找到目标按钮，将指针移动到对应按钮上，最后单击按钮。而用语音控制，则只需要用眼睛寻找到目标按钮后，读出按钮上的字，甚至熟练之后可以抛弃第一步，能够节省大量时间。语音控制和输入可以有效减少用户的操作时间、工作量、学习成本，符合人的原始习惯。这种方法能够有效提升用户体验。

但语音识别存在一定的挑战，语音识别的准确性仍需要改进，口音、方言和小众语言仍然是语音识别所遇到的困难。而且语音输入在公共场合不具有保密性，一些隐私无法通过

语音输入。对于背景声嘈杂的真实环境，语音输入的准确率可能急剧下降，导致交互的失败。目前的语音识别系统进行文本输入时，通常在输入之后，需要自己手工调整部分词句，而且语音输入对于控制的细粒度无法准确表达。特别是对于缩放和移动等命令，语音输入较难实现灵活且准确的量化。语音输入则不需要键盘，由语音识别系统将用户的语音转换为文字。但由于语音识别仍存在一定的挑战，所以常常需要其他的输入方式进行辅助输入，对于识别错误的部分进行修改。

语音识别在技术上存在一定困难，需要通过程序命令设计来减少识别错误。例如，使用简洁的命令，包括选择更多的音节和更少的单词；减少发音相似的单词作为不同的命令。这些方法可以降低语音识别的难度，让用户的语音命令更容易被系统识别，提升语音识别的准确性。同时，建议避免设置一些不可悔改的命令，因为如果用户附近的人意外触发了命令，用户可以轻松撤销操作。

5.4 体感输入

相对于传统的界面交互，体感交互强调利用肢体动作等进行人与产品的交互，通过看得见、摸得着的实体交互设计帮助用户与增强现实系统进行交流。人类生来可以在不借助五感的情况下感知到自身的四肢、关节和肌肉。这个叫作本体感受。通过识别动作和姿态，可以实现依靠自己的身体来进行交互。体感交互的普及目前依赖于几种常见的设备，如捕捉手指运动的 LeapMotion、捕捉身体运动的 Kinect、捕捉肌电信号的 Myo 等。以 Kinect 为典型代表，它不需要人体佩戴设备的体感交互，使用红外摄像头采集人体运动信号，通过算法识别出人体的动作。

从技术的角度来看，体感输入有别于目前主流的按键（键盘鼠标、遥控器等）和触摸（平板、智能手机等）交互方式。在体感交互过程中，用户能根据情境和需求自然地做出相应的动作，而无须思考过多的操作细节。换言之，自然的体感交互削弱了人们对鼠标和键盘的依赖，降低了操控的复杂程度，使用户更专注于动作所表达的语义及交互的内容。在 2020 年新冠肺炎疫情严重的时候，体感交互技术的应用可以让人们不用接触未消毒的键盘或者触摸屏，允许隔空用手势与计算机进行交互，这也降低了病毒感染的可能性。

对于增强现实系统，在不依赖于外界设备的条件下，有若干方式可以实现体感输入。

（1）头戴式增强现实设备，如 Google Glass 等，普遍可以支持手势识别、头部交互等的体感功能。

（2）手持式增强现实设备，如手机、平板等，因为手需要抓握设备，并不方便用于手势交互。这个时候可以探索的是利用头部、眼神等身体部位进行输入。

另外一个思路是增加外部设备，例如上面提到的 Kinect 和 LeapMotion。研究人员提出了基于 Kinect 传感器，利用新型指环作为输入接口的体感识别 Air-Writing（见图 5.7）。设备可通过 3D 手指运动来实现复杂的空间交互，对连续写入空中的整个序列进行识别和连续识别。用户在空中书写字符，就像使用虚构的白板一样。用户可以实时编写大写、小写英文字母以及数字，并且准确率超过 92%。通过 Kinect 跟踪用户手指的运动，提取手指空间位置的变化，由系统重构出手指运动的轨迹。并且当手指到达某个 3D 位置时，用户会从指环以振动的形式接收物理反馈。

图 5.7　以新型指环作为输入方式的体感识别输入设备

另外一个工作探讨了在虚拟环境中使用足部压力分布进行运动类型识别。足底压力的分布由每个鞋底上三个稀疏的传感器检测。系统选择长短时记忆神经网络模型作为分类器，根据压力分布信息来识别用户的运动姿态(见图 5.8)。训练好的分类器直接获取带噪声的稀疏传感器数据，并识别为七种运动姿态(站立、向前/向后行走、奔跑、跳跃、左/右滑动)，而无须手动定义用于对这些姿势进行分类的信号特征。即使存在较大的传感器变化或个体差异，该分类器也能够准确识别运动姿态。结果表明，对于使用不同鞋号的不同用户可以达到接近 80%的精度，而对于使用相同鞋号的用户可以达到 85%的精度。系统提出了一种新颖的方法，将姿势识别的等待时间从 2s 减少到 0.5s，并将准确性提高到 97%以上。这种方法的高精度和快速分类能力可以进一步拓展体感交互在增强现实系统中的应用。

图 5.8　通过足部压力分布识别人体的七种运动姿态

5.5　眼 神 输 入

通过眼神与增强现实系统互动，准确而言是体感输入的一种。用户通过移动目光来移动指针，再凝视或者眨眼来进行确定。在增强现实应用中，眼神作为交互模态的重要性日益凸显。实现眼神输入的前提是准确跟踪眼球的运动。眼动跟踪的基本原理是通过摄像头捕捉眼睛反射的红外光来跟踪人的视线。眼动追踪的最重要目的是改善增强现实应用用户体验。越来越多的头戴式显示器搭载了眼球跟踪设备(见图 5.9)，用于捕捉用户的眼球运动。在手持设备中，前置摄像头也用来捕捉用户的眼球运动。这些改进将帮助增强现实应用的开发者在竞争中脱颖而出。

眼动设备能够了解用户的注意力在三维世界中的分布，带来了大量有关用户注意力的数据源。这是了解用户行为并准确、及时地为用户提供他们想要看到的内容的新前沿技术。一些获得用户同意的专业广告应用程序会跟踪用户在观看网页时的眼动，向广告客户确切显示有多少用户看过广告，注意到他们的品牌推广并消费了关键的营销信息。在可穿戴类

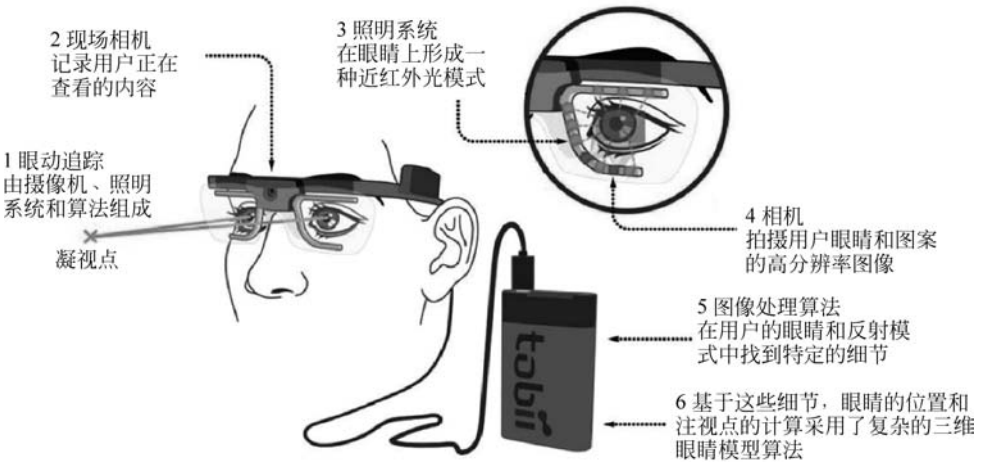


图 5.9　头戴式增强现实设备采用 Tobii 眼球跟踪的示意图

的增强现实设备中，眼动追踪技术可以跟随用户穿越真实街道，并记录用户在真实世界中的信息，特别是对于广告信息的关注程度。广告商、零售商可以相应地优化广告投放位置和商店内的物品布局。

增强现实设备显示效果常常和人的眼睛位置相关，由于并非每个人的眼睛都是一样的，所以为了提供最佳的图像质量，设备需要进行一些微调以补偿眼睛位置的个体差异。一方面是测量眼睛之间的距离(Inter Pupillary Distance，IPD)。IPD 在用户之间差异很大，并且是保持光学清晰度和图像质量的重要因素。使用眼动追踪传感器可以测量并自动调整 IPD，或者可以指导用户达到最佳设置。自动的 IPD 调整可以解决头戴式增强现实设备的障碍之一：无法保持精确的镜头对准。当用户的眼睛花费更少的精力聚焦到关键信息时，这也可以减轻增强现实应用操作过程中的眼睛疲劳程度。高质量的图像将允许用户更长时间地沉浸其中。

当前，眼动追踪的设备达到了一个全新的水平。如果计算机端应用需要实现眼动跟踪，只需要一个眼动捕捉仪。Tobii 是眼动追踪市场的行业领导者之一，在眼动追踪方面拥有数十年的经验，并且可以跨多个行业实现所有类型的应用程序。Tobii 作为“技术开发人员”和“集成商”，将眼动追踪带到多个行业，甚至包括虚拟现实头盔内都可以方便集成眼动跟踪的设备。在头戴式增强现实的硬件上，受限于设备大小，可能采用单目摄像机，通过画面采集去分析人眼运动是可行的方案。在手持设备上，利用前置摄像机去跟踪人眼运动则逐渐成为主流方案之一。

在增强现实技术中，跟踪人眼所关注位置的一个直接受益的技术流程是注视点渲染。注视点渲染旨在依据用户关注的不同区域，采用不同级别的渲染分辨率(见图 5.10)。我们的眼睛在全分辨率下具有狭窄的视野，并且在视网膜中心的外侧出现模糊。以全分辨率渲染整个屏幕会浪费计算系统中的计算资源，因为我们的眼睛无法以全分辨率看到所有内容。注视点渲染能够实现更高分辨率的显示，但不需要设备渲染完整分辨率，可以节省大量 GPU 资源。注视点渲染显示的图形更好地匹配了我们观察对象的自然方式，并带来了许多优点。注视点渲染可以在当前一代的图形处理单元(GPU)上实现 4K 显示，或者在不降低性能的前提下，允许相同的应用程序在成本较低的硬件上运行。

图 5.10　注视点渲染示意图

注视点渲染分为静态注视点渲染和动态注视点渲染。静态注视点渲染集中在固定的区域，无论用户的视线如何，最高分辨率都位于观看设备的视场中心。它通常跟随用户的头部移动，但是如果用户将视线移离观看设备的视场中心，则图像质量会大大降低。动态注视点渲染遵循用户使用眼动追踪的凝视，并在用户视网膜所见之处而不是在任何固定位置渲染清晰的图像。与静态注视点渲染相比，动态方式可为用户提供更好的用户体验和图像质量。由于眼动跟踪能够更好地进行动态注视点渲染，默认接下来所提到的注视点渲染是动态注视点渲染。

实现注视点渲染需要许多不同的硬件和软件紧密集成在一起。具体来说，整个图像渲染链的延迟和同步是关键。眼动追踪算法必须进行高度优化，并在足够好的处理硬件上运行。眼睛跟踪系统将凝视点信息快速传递给系统，告诉 GPU 如何正确渲染图形。此过程必须在几毫秒内发生，否则用户会注意到他们在看的地方与正确渲染的地方之间存在延迟。所有组件必须紧密集成，否则对时滞的敏感性将影响用户体验。

北京大学的研究者提出了一种眼动跟踪的模型 DGaze。这个模型基于卷积神经网络，对动态虚拟场景中的用户注视行为进行分析，用于基于头戴式显示设备中的注视点预测（见图 5.11）。DGaze 首先在自由观看条件下的 5 个动态场景中收集了 43 个用户的眼睛跟踪数据。接下来，DGaze 对数据进行统计分析，并观察到动态对象位置，头部旋转速度和显著区域与用户的注视位置相关。DGaze 结合了对象位置序列、头部速度序列和显著性特征来预测用户的凝视位置。DGaze 不仅可以用于预测实时注视位置，而且可以预测接下来的注视位置，并且可以实现比现有方法更好的性能。在实时预测方面，基于角度距离作为评估指标，DGaze 在动态场景中比以前的方法提高了 22.0%，在静态场景中则提高了 9.5%。研究者将 DGaze 应用到视线渲染和游戏中，并验证模型中每个组件的有效性。

图 5.11　眼动跟踪模型 DGaze 对头戴式显示器中用户的注视点进行预测

5.6 多模态融合

增强现实通常不只有单种模态输入，往往可能有多个模态输入。多模态交互已成为增强现实的研究趋势之一。多模态融合被视为改善虚拟实体与物理实体之间交互的解决方案。因为它基于单模态，又比单模态更加准确。在单模态分析和解释中遇到的困难可以通过将它们集成到多模态系统中来克服，它不仅有利于增强可访问性，而且还带来更多便利。例如，近年来将各种手势识别和语音识别输入引入增强现实中，这在解决增强现实环境中潜在的用户交互限制方面引起了研究者们的极大兴趣。结合手势识别的交互技术为语音识别提供了单独的互补形式。一方面，语音识别的输入可以肯定手势命令，手势可以消除嘈杂环境下的语音识别错误。另一方面，与语音相辅相成的手势只有在与语音一起并可能会注视的情况下才能携带完整的交流信息。HoloLens 的文本输入方法是手势＋指点的混合输入方式。其他的混合方式输入，都是基于被混合的单种输入方式进行相互补充。

西安交通大学的研究者进行了一项实证研究，以结合两种输入机制（滑动输入和按键输入）研究四种输入方法（控制器、头部、手部和混合）的用户偏好和文本输入性能。图 5.12 展示了这四种不同的输入方法。这项研究是对这八种可能组合的首次系统研究。研究结果表明，在文本输入性能和用户体验方面，控制器优于所有其他无需设备的方法。但是，可以根据任务要求以及用户的喜好和身体状况使用无设备单击方法。

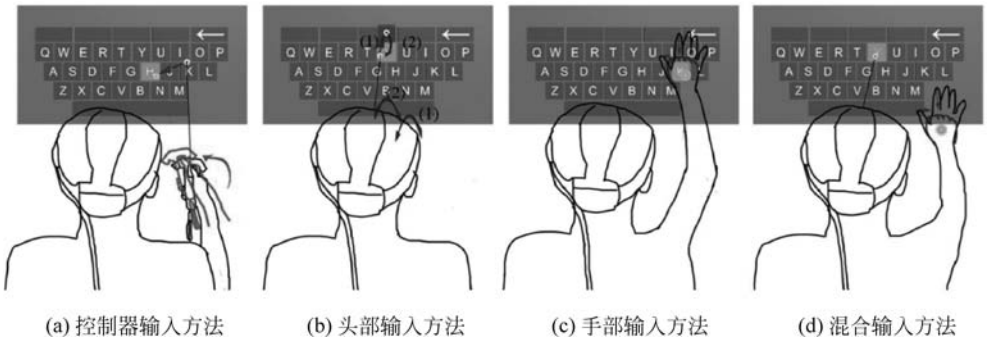


图 5.12　不同的多模态方式用于文本输入

多模态融合系统可以分为两类：特征级融合和语义级融合。特征级融合是在将各种输入形式的信号发送到它们各自的分类器之前，完成功能级别融合。特征级融合被认为是用于集成紧密耦合和同步的输入信号（如信号相互对应的手势识别和语音识别）的一种优秀策略。特征级融合的典型缺点是建模复杂、计算量大且难以训练。通常，特征级融合需要大量的训练数据集。语义级别的融合是在信号从各自的识别器中解释出来之后进行的。语义级别融合适用于集成两个或多个提供补充信息的信号，例如语音识别和手写识别。各个识别器用于独立解释输入信号。可以使用现有的单模态训练数据集来训练那些分类器。因此，输入形式的信号需要彼此具有互补信息，并且时间节点在融合两种不同的输入形式方面起着重要的作用。所识别的输入形式信号的语义表示对于多模态融合是必不可少的，并且相互消除歧义对于解决单一模态的交互错误是必要的。

研究人员探索了一种多模式交互方式，为用户提供了操作虚拟三维物体的方法（见图 5.13）。这种方式结合了眼神和手势跟踪技术，通过二者的结合在虚拟空间中对物体进行选择和操作（包括平移、旋转和缩放等）。通过眼神跟踪实现的注意力机制可能导致操作错误（例如，超过边界、错误选择邻近物体）。因此，结合手势操作，可以更准确地操作物体；而结合眼神操作，可以提升操作的效率。二者相得益彰。

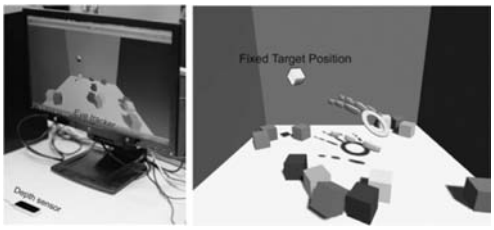


图 5.13　结合手势和眼神跟踪的多模态输入方式在虚拟场景中对物体进行操作

实现多模态交互主要有两个阶段。第一阶段中，要根据交互任务选择各种输入形式；第二阶段是融合所选择的多模态信号。

第一阶段的主要目标是定义可靠和可用的输入形式及其融合。要先确定选择哪种输入形式，如语音识别、手势识别和眼神识别。此阶段的工作是确定单模态输入形式其局限性所引起的问题，以及多模态如何改善或者解决这些问题。输入形式的含义可以根据上下文、任务、用户和时间而变化。输入形式具有非常不同的特征，它们可能没有明显的相似点，而且组合起来也不容易，其中最具挑战性的方面是时间维度。不同的输入形式可能具有不同的时间限制以及不同的信号和语义承受力。例如，手势识别之类的某些输入形式在稀疏、离散的时间点提供信息，而其他模态则生成连续但不像时间那样特定的输出，例如眼神。但这些不同的输入形式常常可以互相弥补。

第二阶段的融合通常是多模态交互系统的关键技术挑战。因为一旦选择了所需的模态，将要解决的一个重要问题是如何将它们组合在一起。为了解决此问题，了解集成模态在增强现实环境中的关系将很有帮助。例如，某些输入形式之间（例如语音和嘴部动作）比其他输入形式之间（例如语音和手势动作）更紧密地联系在一起。通常，尝试不同的输入形式进行组合也是合理的，以使用多种方法执行实际的融合。在技术上，我们应该有每种信号模态的历史记录。通过分析每个输入形式的信号，可以获得其统计特征。然后，可以使用具有提供的统计特性的多通道信号融合，通过多模态融合系统，以有效地合并两个或多个输入形式。

小　　结

本章重点介绍除图像以外的其他增强现实系统的输入，包括键盘、语音、体感、眼神等方式。这些方式将配合图像，作为增强现实系统的输入方式，帮助增强现实系统更好地了解所处的真实世界，更有效地获取用户意图。第 4 章的环境感知技术可以说主要服务于对真实世界的理解，而本章所介绍的多模态输入更多是在增强现实系统与用户交互过程中所用到的方式。

习　　题

1. 简述在之前的其他应用场景下(不局限于增强现实),利用多种模态进行指令输入的体验。

2. 在虚拟现实、增强现实环境下尝试已有的键盘输入、语音输入等方式,评价用户体验,找出不足。

3. 评价各种模态在增强现实场景下的应用价值,以及不同模态组合的互补性。

4. 分析在移动(手机、平板)平台上可以采用的多模态感知技术。

5. 思考是否有现在尚未普及的交互模态,可以应用于增强现实应用的场景中。