E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2022)06-1956-32

论文引用格式:Tao J H, Wu Y C, Yu C, Weng D D, Li G J, Han T, Wang Y T and Liu B. 2022. A survey on multi-modal human-computer interaction. Journal of Image and Graphics,27(06):1956-1987(陶建华,巫英才,喻纯,翁冬冬,李冠君,韩腾,王运涛,刘斌. 2022. 多模态人机交互综述. 中国图象图形学报,27(06):1956-1987)[DOI:10.11834/jig.220151]

多模态人机交互综述

陶建华[1*],巫英才[2],喻纯[3],翁冬冬[4],李冠君[1],韩腾[5],王运涛[3],刘斌[1]

1. 中国科学院自动化研究所,北京　100190;2. 浙江大学,杭州　310058;3. 清华大学,北京　100084;
4. 北京理工大学,北京　100081;5. 中国科学院软件研究所,北京　100190

摘　要:多模态人机交互旨在利用语音、图像、文本、眼动和触觉等多模态信息进行人与计算机之间的信息交换。在生理心理评估、办公教育、军事仿真和医疗康复等领域具有十分广阔的应用前景。本文系统地综述了多模态人机交互的发展现状和新兴方向,深入梳理了大数据可视化交互、基于声场感知的交互、混合现实实物交互、可穿戴交互和人机对话交互的研究进展以及国内外研究进展比较。本文认为拓展新的交互方式、设计高效的各模态交互组合、构建小型化交互设备、跨设备分布式交互、提升开放环境下交互算法的鲁棒性等是多模态人机交互的未来研究趋势。

关键词:多模态人机交互;大数据可视化交互;声场感知交互;实物交互;可穿戴交互;人机对话交互

A survey on multi-modal human-computer interaction

Tao Jianhua[1*], Wu Yingcai[2], Yu Chun[3], Weng Dongdong[4], Li Guanjun[1],
Han Teng[5], Wang Yuntao[3], Liu Bin[1]

1. Institute of Automation, Chinese Academy of Sciences, Beijing　100190, China;
2. Zhejiang University, Hangzhou　310058, China; 3. Tsinghua University, Beijing　100084, China;
4. Beijing Institute of Technology, Beijing　100081, China; 5. Institute of Software, Chinese Academy of Sciences, Beijing　100190, China

Abstract: Benefiting from the development of the Internet of things, human-computer interaction devices have been widely used in people's daily life. Human-computer interaction is no longer limited to the input and output modes of a single sensory channel (vision, touch, hearing, smell and taste). Multi-modal human-computer interaction aims to exchange information between human and computer by using multi-modal information such as speech, image, text, eye movement and touch. Multi-modal human-computer interaction includes multi-modal information input from human to computer and multi-modal information presentation from computer to human and it is a comprehensive discipline closely related to cognitive psychology, ergonomics, multimedia technology and virtual reality technology. At present, multi-modal human-computer interaction and various kinds of academic and technology in the field of image and graphics are more and more closely combined. In the era of big data and artificial intelligence, multi-modal human-computer interaction technology, as the technical carrier of human-machine-thing, is closely related to the development of image and graphics, artificial intelligence, emotional computing, physiological and psychological assessment, Internet big data, office education, medical rehabilitation and other fields. The research on multi-modal human-computer interaction first appeared in the 1990 s, and a number of works proposed an interactive method combining speech and gesture. In recent years, the emergence of immersive visualization provides a new multi-modal interactive interface for human-computer interaction: an immersive environment that integrates vis-

收稿日期:2022-02-26;**修回日期**:2022-03-23;**预印本日期**:2022-03-30

*通信作者:陶建华　jhtao@nlpr.ia.ac.cn

ual, auditory, tactile and other sensory channels. Visualization is an important scientific technology for data analysis and exploration. It converts abstract data into graphical representations and facilitates analytical reasoning through interactive interfaces. In today's data explosion, visualization transforms complex big data into easy-to-understand content, improving people's ability to understand and explore data. The traditional interactive interface can only support a flat visual design, including data mapping channels and data interaction methods, and cannot meet the analysis needs in the context of the big data area. In the area of big data, data visualization will have problems such as limited presentation space, abstract data expression, and data occlusion. The emergence of immersive visualization provides a broad presentation space for high-dimensional big data visualization, integrating multi-sensing channels and multi-modalities. Interaction allows users to interact with data naturally and in parallel using multiple channels. The interaction technology based on sound field perception can be divided into three types according to the working principle: measure and identify the acoustic characteristics of a specific space, passage or the change of the acoustic characteristics caused by the action; use the sound wave measurement of the microphone array to achieve sound source localization, the sound source can emit specific carrier audio to improve the positioning accuracy and robustness; the machine learning algorithm recognizes the sound from a specific scene, environment or human body. The technical solution includes a single method based on sound field perception and a sensor fusion solution. In the physical interaction system, the user interacts with the virtual environment by using the physical objects existing in the real environment. In recent years, the integration of physical interaction interface technology into virtual reality and augmented reality has become a mainstream direction in this field, and the concept of "physical mixed reality" has gradually formed, which is also the conceptual basis of passive haptics. The haptics of physical interaction can be divided into three ways: static passive haptics; passive haptics with feedback and active force haptics. Since active haptic devices are relatively expensive, there are few current researches, and the main research directions are still static passive haptics and encounter-type haptics. Regarding the mixed reality interaction mode of passive haptics, the current research levels of various countries and institutions in the world are not very different, but there is a slight emphasis. Wearable interaction is mainly divided into research on gesture interaction and touch interaction mainly in the form of wristbands, skin electronic technology and interaction design. Gesture input is considered to be one of the core contents of "natural human-machine interface", and it is also suitable for exploring the input methods of wearable devices. The key to realizing gesture input lies in sensing technology. At present, in the field of human-computer interaction, the sensing technology for gesture recognition based on infrared light, motion sensor, electromagnetic, capacitive, ultrasonic, camera and biological signals has been deeply studied. As the natural interface between people and the outside world, the skin has been initially used to explore its role in information interaction, and its applications in several aspects have demonstrated its advantages. The human-computer dialogue interaction process involves multiple modules such as speech recognition, emotion recognition, dialogue system, and speech synthesis. First, the user's speech is converted into corresponding text and emotion labels through speech recognition and emotion recognition modules. The dialogue system is then used to understand what the user is saying and generate dialogue responses. Finally, the speech synthesis module converts the dialogue responses into speech to interact with the user. How to effectively integrate information of different modalities in the human-computer interaction system and improve the quality of human-computer interaction is also worthy of attention. Multi-modal fusion methods can be divided into three types: feature layer fusion methods, decision layer fusion methods, and hybrid fusion methods. The feature layer fusion method maps the features extracted from multiple modalities into a feature vector through a certain transformation and then sends it to the classification model to obtain the final decision. The decision-level fusion method combines the decisions obtained from different modal information to obtain the final decision. The hybrid fusion method adopts both the feature layer fusion method and the decision layer fusion method. This paper systematically reviews the development status and emerging directions of multi-modal human-computer interaction, and thoroughly combs the research progress of big data visualization interaction, interaction based on sound field perception, near-eye display entity interaction, wearable interaction, and human-computer dialogue interaction. This article believes that expanding new interaction methods, designing efficient interaction combinations of various modalities, building miniaturized interactive devices, cross-device distributed interaction, and improving the robustness of interactive algorithms in open environments are the future works of multi-modal human-computer interaction.

Key words: multi-modal human-computer interaction; big data visualization interaction; sound field perception interaction; entity interaction; wearable interaction; human-computer dialogue interaction

0 引 言

受益于物联网的发展,人机交互设备在人们的日常生活中得到了广泛应用。近年来,计算机视觉、手势识别和人工智能等技术蓬勃发展,头戴式设备、显示屏和传感器等硬件技术取得了明显的进步,人机交互不再局限于单一感知通道(视觉、触觉、听觉、嗅觉和味觉)的输入输出模态(Bourguet,2003)。

多模态人机交互旨在利用语音、图像、文本、眼动和触觉等多模态信息进行人与计算机之间的信息交换。其中包括人到计算机的多模态信息输入与计算机到人的多模态信息呈现,是与认知心理学、人机工程学、多媒体技术和虚拟现实技术等密切相关的综合学科。目前,多模态人机交互与图像图形领域中的各类学术和技术联合得越来越紧密。多模态人机交互技术作为人—机—物的技术载体,在大数据与人工智能时代,其学术和技术发展前沿与图像图形学、人工智能、情感计算、生理心理评估、互联网大数据、办公教育和医疗康复等领域发展息息相关。多模态人机交互研究最早出现在 20 世纪 90 年代,多项工作提出了将语音和手势融合在一起的交互方法(Pavlovic 等,1997;Ando 等,1994;Cassell 等,1994)。近几年,沉浸式可视化(Jansen 等,2014)的出现为人机交互提供了一个新的多模态交互界面:一个融合了视觉、听觉和触觉等多个感知通道的沉浸式环境。

在学术界,多模态人机交互的学术成果在 IEEE-TPAMI(IEEE Transactions on Pattern Analysis and Machine Intelligence)、IEEE-TIP(IEEE Transaction on Image Processing)、IEEE-TASLP(IEEE/ACM Transactions on Audio, Speech and Language Processing)、IEEE-TNNLS(IEEE Transactions on Neural Networks and Learning Systems)、ACM-TOCHI(ACM Transactions on Computer-Human Interaction)等国际期刊和 CHI(Computer-Human Interaction)、UbiComp(Ubiquitous computing)、CSCW(ACM Conference on Computer-Supported Cooperative Work and Social Computing)等国际会议呈现稳步增长,创新成果层出不穷。

在产业界,语音、人脸和手势等新型交互的应用从噱头转趋理性,聚焦于车载、直播等特定场景。触屏搭配一种新模态的交互方式,是当前多模态交互产品落地的主要形态。增强现实等新型输出/显示模态的技术逐渐成为未来多模态人机交互产品新的主要场景。

各国政府高度重视多模态人机交互。在“十三五”期间,我国设立多项重大重点项目支持多模态人机交互方向的研究。例如,国家重点研发计划项目“基于云计算的移动办公智能交互技术与系统”、“多模态自然交互的虚实融合开放式实验教学环境”等。美国海军开始构建下一代舰艇多模态人机交互模式,采用全息化的指挥模式,通过佩戴视觉和触觉传感器对舰船进行控制。英国海军公布的 T2050 未来水面舰艇概念,以多模态人机交互的方式,有效提高工作效率。

本文旨在综述多模态人机交互的最新进展,帮助初学者快速了解和熟悉多模态人机交互领域;对多模态人机交互方式进行分类整理,帮助该领域的研究者更好地理解多模态人机交互中的各种技术;对多模态人机交互领域面临的机遇和挑战进行梳理,启发相关研究者做出更有价值的多模态人机交互工作。

本文将从多模态信息输入与多模态信息输出两方面对多模态交互技术进行综述。其中,多模态信息输入过程涉及可穿戴交互技术以及基于声场感知的输入交互技术。多模态信息呈现过程涉及大数据可视化交互技术、混合现实交互技术以及人机对话交互技术。下面分别从大数据可视化交互、基于声场感知的交互、混合现实实物交互、可穿戴交互和人机对话交互 5 个维度介绍多模态人机交互的研究进展。内容框架如图 1 所示。

1 国际研究现状

1.1 大数据可视化交互

可视化是一种数据分析和探索的重要科学技术(叶帅男 等,2021),将抽象数据转换成图形化表征,

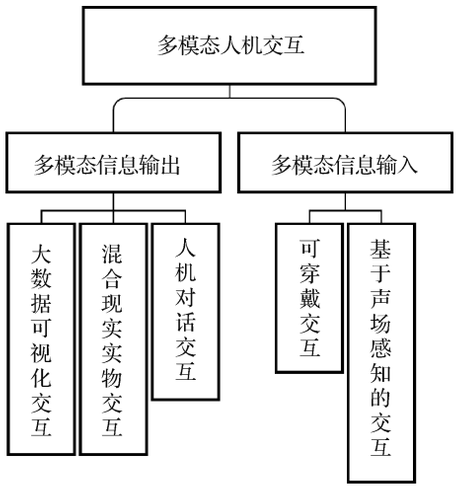


图 1　本文内容框架

Fig. 1　The architecture of this paper

通过交互界面促进分析推理，在城市规划（Deng 等，2021）、医疗诊断（Park 等，2021）和运动训练（Chu 等，2022）等领域起着关键作用。在数据爆炸的今天，可视化将纷繁复杂的大数据转换为通俗易懂的内容，提升了人们理解数据和探索数据的能力。

传统的可视化交互设备，无论加载何种可视化系统，皆以 2 维显示屏、键盘和鼠标三者构成为主，通过键盘鼠标进行点击、拖拽、框选和移动等交互对可视化内容进行探索。然而，此交互界面只能支持平面式的可视化设计，包括数据映射通道、数据交互方式，无法满足大数据时代背景下的分析需求。

数据可视化在大数据时代下会产生呈现空间有限、数据表达抽象和数据遮挡等问题，沉浸式可视化的出现为高维度的大数据可视化提供了广阔的呈现空间，综合了多感知通道的多模态交互使用户可以利用多通道自然而并行地与数据交互。

1.1.1　大数据可视化设计

如何可视化复杂结构的海量数据依旧是一个挑战，尤其是具有 3 维空间信息的数据。传统的平面式呈现将视觉通道和视觉反馈局限于 2 维空间中（Ma 等，2014），同时也限制了设计空间。沉浸式设备的发展释放了用户的立体视觉，研究者们开始发掘3 维交互空间在可视化中的潜力。

人们对 3 维的视觉感知来自于双目视差、遮挡和相对大小等深度提示（Renner 等，2013）。一方面，用户能够轻易识别 3 维物体的形态；另一方面，3 维中的视角倾斜会使 2 维平面图形产生形变，使用户难以识别（Munzner，2014）。因此，如何在 3 维环境中进行有效的可视化设计是大数据可视化交互领域的研究热点之一。

点是可视化中的重要标记。在 2 维平面中，通常可以采用点的位置、大小和颜色等视觉通道编码数据的不同属性。在 3 维环境中，Kraus 等人（2020）通过用户实验发现相比于 2 维平面上的散点图，用户可以在虚拟现实环境下更加有效地识别 3 维散点图中的聚类。Alper 等人（2011）提出了一种在 3 维环境中对图数据结构进行可视化的方法。该技术利用立体深度，通过将用户感兴趣的区域投影到更靠近用户视线的平面上进行突出显示。然而，上述可视化方法占据了 3 维位置的视觉通道，因此不能编码点在 3 维环境中的位置。为了解决上述问题，Krekhov 和 Krüger（2019）以及 Krekhov 等人（2020）提出了 Deadeye 技术，通过分裂呈现的方法对点进行突出显示。如图 2 所示，分裂呈现技术根据对每只眼睛呈现不同的刺激，将需要高亮的点在一只眼中显示。通过这种技术，需要高亮的点可以立即被视觉系统检测到。

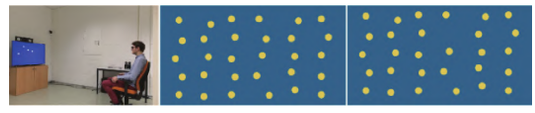


图 2　分裂呈现技术效果图（Krekhov 等，2020）

Fig. 2　Effect of split rendering technology (Krekhov et al., 2020)

线等视觉元素广泛应用于时空数据可视化中。然而传统的 2 维流图将同一区域不同时间的流动投影至一片区域中，造成不同时间流动情况相互覆盖。时空立方体是一种在 3 维环境下对时空数据进行直观可视化的方法。时空立方体采用水平方向上的两个维度编码位置信息，采用垂直方向上的维度编码时间信息。Ssin 等人（2019）提出了一种基于时空立方体对轨迹数据进行可视化的技术 GeoGate。GeoGate 是一种增强现实环境下的可视化系统。该系统扩展了时空立方体，并采用一个环形用户界面来探索多个位置数据集中实物之间的相关性。Filho 等人（2019）提出了一种虚拟现实环境下的时空数据可视化系统。该工作使用时空立方体构建虚拟现实环境下的原型系统，将多维数据集与用户桌面的虚拟表示相结合。在展示地理流动数据的

场景中,Yang 等人(2019)通过增加高度,将 2D 地图中的流分开,通过实验得出将流按照不同高度展示可以提高用户辨别地图中流的准确率。

图可视化是信息可视化中的一个重要领域。传统的 2 维图布局会在数据量增大时出现严重的遮挡问题,为了解决此问题,Kwon 等人(2016)提出了沉浸式环境下的图可视化的布局、渲染和交互技术的设计,提高了人们对大数量级图可视化的探索分析能力,如图 3 所示。

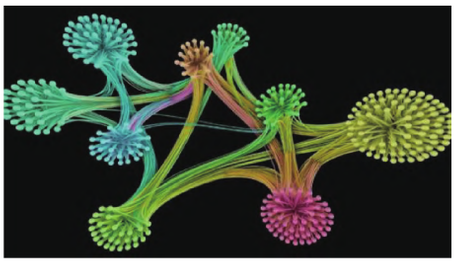


图 3　沉浸式图可视化(Kwon 等,2016)

Fig. 3　Immersive graph visualization (Kwon et al., 2016)

1.1.2　非视觉感知的交互辅助

非视觉感知包括听觉、触觉、嗅觉与味觉。这些感知在日常生活中为人们提供了大量的信息,例如方位、声音和温度等,并与视觉一起帮助人们感知与理解周围的事物。近年来,多模态硬件技术愈发成熟,用以产生或模拟非视觉感知的设备逐步地小型化与商业化,这促使大数据可视化交互领域开始研究非视觉的交互方式。这种数据交互方法将用户沉浸在数据中,并在视觉感知外提供听觉、触觉等感知通道,提升用户的参与感与沉浸感,让用户感知在单一视觉通道上难以被发现的细节和模式。

在非视觉感知中,听觉是最容易实现的感知通道。通过物体发出的立体声,用户能够轻易辨识其所在的方位、远近等信息(Siu 等,2020);而语音则可高效地传递描述数据的语义信息(Kong 等,2019)。声音的音调、音色、音量以及听者所在的位置都可作为数据映射的通道并用以编码类别以及连续的数据类型,例如 Franklin 和 Roberts(2003)将饼图中的类别信息、占比转化为各类型的声音;Xi 和 Kelley(2015)则提出了利用声音分析时序数据的工具。

触觉感知能够为用户提供物体纹理、温度和振动幅度等类别或连续的信息。利用振幅的大小,Prouzeau 等人(2019)将 3 维散点图中点云的密度映射为不同振幅的等级,提升了用户发现点云中心高或低密度的区域的感知能力。此外,数据物理化则是将抽象数据转化为可触摸实物的方法,通过让用户与实物触摸而不仅仅是观看来提升探索数据的效率,例如柱状图(Hu,2015)、节点链接图(Dragicevic 等,2021)等。

嗅觉与味觉具有易于记忆和识别的优势。利用各种气味所提供的类别信息以及气流流速、温度等连续信息,嗅觉与味觉同样能够编码离散与连续的数据。例如 viScent(Patnaik 等,2019)提出了不同气味与数据类型的映射空间以将数据编码为不同的气味。

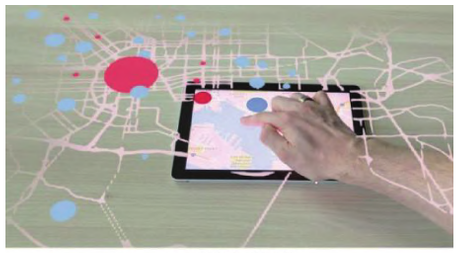
非视觉感知作为视觉感知的补充,能够提升用户分析理解数据的效率,例如在分析大量或高密度分布的数据可视化时,让用户感知视野之外或被遮挡的数据信息。另外,对于部分无法获取大数据可视化中视觉信息的人群而言,非视觉感知的交互能将可视化中的信息转化成非视觉信息传达给他们。然而,这些感知的使用往往带来额外的疲劳感,例如长时间触摸所导致的手臂疲劳,进而降低分析的时长。同时如何将高维、多变量等复杂数据进行非视觉感知的编码与设计仍尚待研究。

1.1.3　多模态交互设计

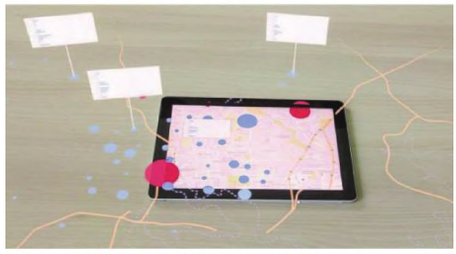
在大数据可视化交互领域,除了可视化设计,现有的研究重点还集中在探索更加自然直观的交互方式,以提升人们在 3 维空间对大数据可视化的操作效率。多模态交互结合单一模态的优点,充分发挥了人们对各个感知通道传达信息的高度接收与处理能力,增强用户对交互行为的理解,提高对大数据可视化的探索与分析效率。

1)基于接触的交互。以智能手机、平板为主的移动设备为可视化交互提供了高清晰度的画面和高精度的交互。基于接触的交互支持用户直接通过手部或者手持传感器触碰可视化标记,传递对数据的交互表达,这类自然的交互方式的操作精度较高,并且能够增强用户在探索大数据可视化时对信息的理解。如图 4 所示,Langner 等人(2021)通过平板触控的方式准确地选择可视化图表以更新 HoloLens 中所见的增强内容。

除了增强现实环境中基于触屏的交互方式外,接触式交互在虚拟现实中也很常见。例如,如图 5



(a) 地图可视化



(b) 神经系统可视化

图 4　通过平板触控的交互(Langner 等,2021)

Fig. 4　Interaction through tablet touch(Langner et al. ,2021)

((a)map visualization;(b)nervous system visualization)

所示,Usher 等人(2018)的 VR(virtual reality)系统可以通过跟踪用户手部动作来捕获用户勾勒出来的脑神经路径。与数据交互后产生的触觉反馈可以提升用户交互的直观性和沉浸感。研究者探索了虚实物体结合的交互方式,将真实物体作为虚拟标记在真实世界中的参照物给用户触碰来提升交互的精确性。例如,研究者将沉浸空间中的地图或数据点投影等 2 维可视化平面视为如图 5 所示的虚拟桌面并将其映射至真实桌面(Wagner 等,2021),用户可以直接点击桌面来操作对应数据。此外,Cordeil 等人(2020)使用 3 个滑块轴将数据坐标轴实物化,用户可以通过操作滑块的位置来精准地选择轴空间内的数据。

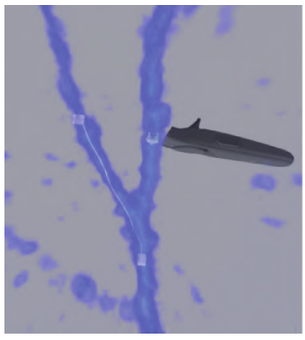


图 5　通过手部跟踪捕获勾勒的路径(Usher 等,2018)

Fig. 5　Capture the outlined path through hand tracking(Usher et al. ,2018)

2)基于手势的交互。动作识别和传感器技术的发展让基于手势的交互逐渐成为常用的交互方式之一。基于手势的交互使用可跟踪设备或捕捉用户手指的移动来捕捉手部动作,帮助用户完成对数据的操控(Büschel 等,2018)。一种常见的手势交互技术是光线投射的目标指向,用户可以使用手柄等装置射出的光线来选择与光线相交的最近对象。为了增加这类交互方式的精确度,RayCursor(Baloup 等,2019)增加了如图 6 所示的沿投射光线方向的红色的控制光标来避免被遮挡散点的选择。此外,FiberClay(Hurter 等,2019)支持用户操控手柄射出的射线来完成对轨迹的筛选,如图 7 所示。

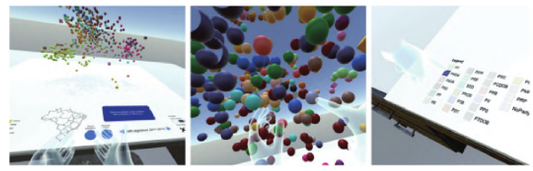


图 6　虚拟桌面示意图(Wagner 等,2021)

Fig. 6　Schematic diagram of VirtualDesk(Wagner et al. ,2021)

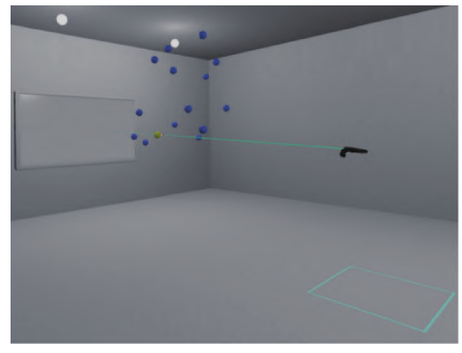
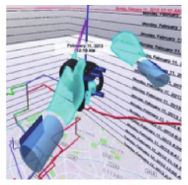


图 7　基于光线投射的交互设计(Baloup 等,2019)

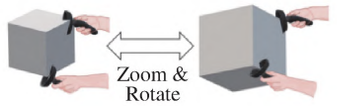
Fig. 7　Interactive design based on ray casting (Baloup et al. , 2019)

除了光线投射技术的指向隐喻,其他诸如抓取、拖动等的隐喻也都有相关研究。如图 8(a)所示,Wagner 等人(2021)采用了虚拟手的隐喻,设计抓取和拉伸等动作完成对时空轨迹可视化的移动、缩放和选择等操作;Yang 等人(2021a)利用双手合拢与展开的手势实现了散点图的缩放操作,如图 8(b)所示;TiltMap(Yang 等,2021b)通过改变手柄的倾斜角度来实现如图 8(c)所示的对地图可视化的不同视图之间的切换。这些交互方式通过直观的手势隐喻,帮助用户减轻了许多交互负担。

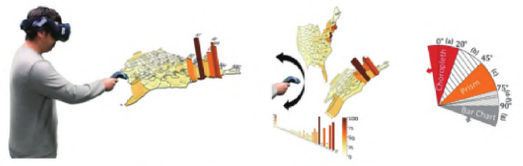
3)基于注视的交互。利用用户的视线信息进行注视交互也是探索大数据可视化时一种常见的交



(a) 用于抓取、拉伸等交互行为的虚拟手隐喻



(b) 双手合拢展开实现可视化的缩放交互



(c) 利用手柄的倾斜改变视图

图 8　3 种基于手势隐喻的交互设计
(Wagner 等,2021;Yang 等,2021a,b)

Fig. 8　Three interaction designs based on gesture metaphor (Wagner et al., 2021; Yang et al., 2021a, b) ((a) virtual hand metaphor for grasping, stretching, and other interactions; (b) scaling interaction for visualizations with pinch gesture; (c) changing views with the tilt of the controller)

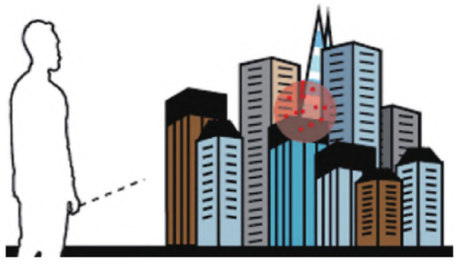


图 9　通过眼动追踪技术完成目标选择的交互设计
(Sidenmark 等,2020)

Fig. 9　Interaction design of target selection through eye tracking technology (Sidenmark et al., 2020)

互模态。基于注视的交互通过眼动追踪技术捕捉用户的视线焦点,从而理解用户视线中传递的信息,例如当前关注的内容,或者用户的心理状态等。更进一步,系统可以基于这些信息完成交互,例如高亮用户关注的内容(Kwok 等,2019)。Sidenmark 等人(2020)使用该技术辅助用户如图 9 所示在虚拟 3 维场景中选择一些被遮挡的物体:用户注视物体轮廓上的圆点,并使用视线控制圆点在物体未被遮挡的轮廓线上移动,从而精准地选中被部分遮挡的物体。Alghofaili 等人(2019)则使用长短时记忆网络(long short-term memory, LSTM)模型对用户眼动数据进行异常检测,从而判断用户是否适应当前的虚拟环境,并在用户迷失时给予辅助反馈。

4)基于移动导航的交互。移动导航也是探索呈现在虚拟的广阔 3 维场景里的大数据可视化中一个重要的交互模态。然而虚拟空间会出现与物理空间不匹配的情况,影响用户在虚拟空间中达到最佳观察点,降低探索能力。为此,交互式的移动导航可以辅助用户移动到最佳的观察点,甚至同时规避空间感知不一致性带来的生理不适。Abtahi 等人(2019a)通过建立 3 个层级的真实速度到虚拟速度的映射,便于用户在狭小的真实空间内遍历数据可视化呈现空间。此外,虚拟移动技术更进一步地拓宽了遍历虚拟空间的可能性。此类技术包括指定位置进行直接传送(Funk 等,2019)、使用 3 维缩略图进行传送(Yang 等,2021a)以及使用手柄控制飞行动作(Drogemuller 等,2018)等。

1.2　基于声场感知的交互

基于声场感知的交互技术按照工作原理可分为以下 3 种:1)测量并识别特定空间、通路的声音频响特性或动作导致的声音频响特性变化;2)使用麦克风组/阵列的声波测距(角)实现声源的定位,可通过发声体发出特定载波音频提升定位精度与鲁棒性;3)机器学习算法识别特定场景、环境或者人体发出的声音。技术方案包括单一基于声场感知的方法与传感器融合的方案。

本节从基于声场感知的动作识别、基于声源定位的交互技术、基于副语音信息的语音交互增强以及普适设备上的音频感知与识别 4 个方面综述国际上基于声场感知的交互技术。

1.2.1　基于声场感知的动作识别

基于声场感知实现不同手势与动作的识别是人机交互的热点研究内容,基于手势或者姿态带来声场变化的基础,实现手势或动作的识别。比如使用耳机上的麦克风识别摘戴耳机是最直观的手势识别,Röddiger 等人(2021)利用内耳麦克风识别出了中耳内鼓膜张肌的收缩等用于交互。对于双手手势的识别,很多研究者会增加扬声器来构建设备周围的声场,通过分析麦克风接受到的信号变化来识别相应的手势。对于笔记本电脑、屏幕等固定设备,研究者使用声场识别出了手在空中的挥动、停留等手势(Ruan 等,2016;Gupta 等,2012)。而手表和腕带等可穿戴设备上的应用则更加广泛,Han 等人(2017)通过手表上的特殊排布的麦克风阵列识别

了手腕的转动、拍手臂、不同位置打响指等手势，BemBand(Iravantchi 等,2019)利用腕带上超声波信号完成了对于手掌姿态、竖拇指等手势的识别。除此之外,很多研究将声音信号与陀螺仪的运动信号结合以实现更加精细的动作识别,早期 Ward 等人(2005)利用两个腕带上的麦克风和陀螺仪进行过简单手势识别的探索。而近几年传感器精度和质量的逐步提升,更多相关研究提高了手势识别的准确度与精度,FingerSound(Zhang 等,2017a)与 FingerPing(Zhang 等,2018)均识别拇指在其他手指上的点击与捏合动作,且 FingerPing 利用了不同手势下的共振信息,减少了对于陀螺仪的依赖,TapSkin(Zhang 等,2016)识别出了手表附近皮肤上的点击等更精细的手势交互动作。除了手势动作本身,部分研究探索了用户在与其他物品交互时的行为和手势,Acustico(Gong 等,2020)利用腕带上贴近桌面的麦克风识别出了用户在桌面上点击的 2 维位置用于交互输入,Pentelligence(Schrapel 等,2018)和 WritingHacker(Yu 等,2016)利用笔上麦克风较准确地还原出用户书写的字迹,而 Ono 等人(2013)利用玩具上的麦克风识别出了用户的触摸位置。

1.2.2　基于声源定位的交互技术

声源定位通常依赖于精确的距离测量。通过不同的声学测距方法,可以得到声源与麦克风的距离;再通过三角定位法,即可得到声源的位置。声学测距的常用方法包括基于多普勒效应、基于相关和基于相位的测距方法,此外在雷达系统中广泛应用的调频连续波(frequency modulated continuous wave, FMCW)也在近些年应用于声学测距。基于以上声学测距技术,可以实现手势识别、设备追踪等交互技术。

基于多普勒效应,通过频率变化来进行精确的距离计算,AAMouse(Yun 等,2015)实现了中位数误差 1.4 cm 的精确追踪,通过追踪手中移动设备的位置,实现了鼠标的功能。基于相关来计算到达时间差,BeepBeep(Peng 等,2007)使用线性调频信号和两路感知技术实现了设备间厘米级精度的距离测量。Tracko(Jin 等,2015)基于 BeepBeep 中提出的算法,融合 BLE(bluetooth low energy)和 IMU(inertial measurement unit),实现了设备间的 3 维空间感知。基于手指、手掌运动导致回波相位的改变,LLAP(Wang 等,2016)实现了 4.6 mm 的 2 维追踪精度,实现了不需要佩戴额外设备的手势追踪。SoundTrack(Zhang 等,2017b)应用类似的技术但将感知范围扩展到了 3 维空间,利用内置扬声器的指环和有麦克风阵列的智能手表实现了对手指的连续追踪。基于 FMCW 技术,CAT(Mao 等,2016)通过结合多普勒效应和 IMU 实现了 8～9 mm 的追踪精度。MilliSonic (Wang 和 Gollakota, 2019)利用 FMCW 中的相位信息计算距离,实现了基于智能手机与四麦克风阵列的原型,达到了 2.6 mm 的 3D 精确度,进一步提高了对智能设备的追踪能力。

除了被追踪设备作为声源主动发声,还可以利用被追踪物体的回波来进行定位。FingerIO(Nandakumar 等,2016)应用正交频分复用(orthogonal frequency division multiplexing, OFDM)技术来追踪手指的回波,不需要在手指上佩戴其他的传感器,实现了平均精度为 8 mm 的 2 维手指追踪。Mao 等人(2019)利用身体和手部的反射信号,实现了一个基于 RNN(recurrent neural network)的房间尺度的手部追踪系统。该系统通过把基于 2D MUSIC(multiple signal classification)方法得到的数据输入到 RNN 中来获取传播距离和到达角信息,在 4.5 m 范围内达到了 1.2～3.7 cm 的追踪精度。

1.2.3　基于副语音信息的语音交互增强

近年来有许多研究者研究了利用“言语中的非言语信息”来加强语音互动。Goto 等人(2002)提出利用语音过程中的用户在元音处的短暂停顿自动显示候选短语辅助用户记忆,并提出了利用用户有意控制的音高移位切换语音输入模式(Goto 等,2003),以及利用语音中的停顿和音高区分连续对话中的人人对话和人机对话(Goto 等,2004)。Kitayama 等人(2003)提出了利用自然语音交互中的口语现象和停顿进行噪音鲁棒的端点检测和免唤醒。Kobayashi 和 Fujie(2013)研究了人—机器人对话中的副语言协议。Maekawa (2004)与 Fujie 等人(2003)讨论了副语言产生和感知的原理。Fujie 等人(2004)研究了利用副语言信息改进对话系统。Harada 等人(2006,2009)研究了利用元音质量、音量和音高等属性的光标控制。House 等人(2009)将这一思想延续到利用连续声音特征控制 3 维机械臂。Igarashi 和 Hughes(2001)研究了利用非言语信息的连续语音控制和速率的参数。

1.2.4　普适设备上的音频感知与识别

近年来,普适音频设备不断普及,产业界对于普

适音频设备不断投入,众多研究者致力于研究普适设备上的音频感知与识别。普适音频设备对于音频数据的实时性采集使得其在声音实时分类事件上具有优势,如 Rossi 等人(2013)提出了利用智能手机麦克风实时进行环境声音识别的系统 AmbientSense。普适音频设备的声音感知还常常用在健康与生理感知领域,用以捕捉、推断用户的生理信息。如 Thomaz 等人(2015)提出利用腕部音频设备捕捉环境声音,进行识别后推断用户饮食活动的方法,帮助用户进行饮食自我监测。Amoh 和 Odame(2015)提出利用可穿戴声学传感器结合卷积神经网络检测咳嗽的技术。与环境的声音检测类似,对于更广义上的用户行为,Lu 等人(2009)利用手机麦克风对人当前活动(开车、乘坐公交车等)的识别进行了探索。商业产品或应用也快速发展与成熟,其中最具有代表性的是苹果手机手表上的环境音感知(咳嗽、报警等)。

1.3 混合现实实物交互

通过真实物体实现与虚拟对象进行交互的方法称为“实物交互界面”(Ishii 和 Ullmer,1997)。在实物交互系统中,用户通过使用在真实环境中存在的实物对象与虚拟环境进行交互,由于用户对实物本身的各种特性(如形状、重量)非常熟悉,可以使得交互的过程更为精准和高效(Zhou 等,2008)。近年来,将实物交互界面技术融入虚拟现实和增强现实已成为本领域的一个主流方向,并逐渐形成了“实物混合现实”的概念,这也正是被动力触觉的概念基础。2017 年,Zhao 等人(2017)将实物交互的触觉分为 3 种方式:1)静态的被动力触觉;2)具有反馈的被动力触觉(即相遇型触觉);3)主动的力触觉。由于主动力触觉装置比较昂贵,目前的研究很少,主要研究方向仍是静态的被动力触觉和相遇型触觉。关于被动力触觉的混合现实交互方式,目前国际上各个国家和机构的研究水平差别不大,但略有侧重。

1.3.1 静态的被动力触觉

在静态的被动力触觉方面,加拿大多伦多大学和美国芝加哥大学等团队曾提出过 Thors Hammer(Heo 等,2018)以及 PHANTOM(Massie 和 Salisbury,1994)两种比较具有代表性的研究。如图 10 所示,通过 1∶1 制作的物理实物道具提供逼真的动觉和触觉反馈,提高用户的触摸感受以及操作能力,并且可以通过对实物的触摸来对虚拟对象进行操作。静态的被动力触觉是在混合现实环境中实现触觉交互的一种早期探索,但这些刚性道具在形状上往往和虚拟道具不匹配,或者是道具数量有限,不能满足交互的需求。因此,可变换的被动力触觉便应运而生。加拿大多伦多大学的 Araujo 等人(2016)提出了 Snake Charmer,可以动态地改变交互对象的纹理特征和材质信息,在虚拟环境中渲染不同的对象时仍能够保持触觉和视觉的一致性(Lee 等,2006)。

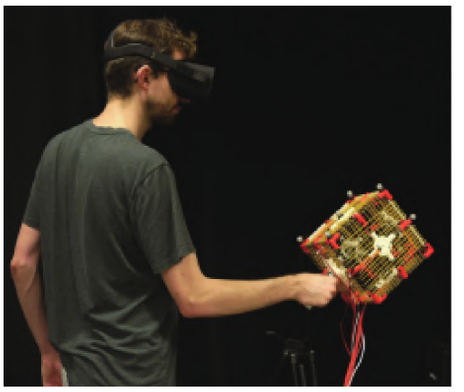


图 10 虚拟现实中的触觉反馈(Heo 等,2018)

Fig. 10 Haptics feedback in virtual reality(Heo et al.,2018)

1.3.2 相遇型被动力触觉及 3 种触觉设备

早在 1993 年,McNeely(1993)就提出机器人图形(robotic graphics)的概念,他认为触觉输出具有极大的价值,并建议使用机械臂或者机器人作为形状载体,动态地提供物理反馈。如今,这种方式已用于混合现实环境中,并有了深远的进步。对于有反馈的被动力触觉系统,从交互道具角度,有反馈的被动力触觉系统的交互载体主要有穿戴式、手持式和机器人式 3 大类(Huang 等,2020a)。与目前市场上主流的交互方式——如 HTC Vive 和 Oculus Quest 的交互手柄相比,基于被动力触觉的混合现实交互方式可以让用户在混合现实场景中更真实地操作物体,并提供力反馈。

1)可穿戴式触觉反馈设备。可穿戴式触觉反馈设备通过触觉手套、触觉服饰等方式,直接将机械系统产生的力反馈或者电反馈施加在用户的手部或身上,直观地进行被动力反馈触觉。美国斯坦福大学的 Choi 等人(2016)提出的 Wolverine 是一个典型的例子。Wolverine 通过低成本和轻量级的设备,可以直接在拇指和 3 根手指之间产生力,以模拟垫式握持式物体,比如抓握茶杯和球。在低功耗的情况

下能反馈超过 100 N 的反馈力。但是,这些可穿戴设备的缺点是,用户在混合现实环境中必须要时刻穿戴着反馈装置,有一定不适感,并且难以实现裸手交互。

2)手持式触觉设备。手持式触觉设备是通过用户单手或双手抓握指定的物体,从而对用户实现力反馈,具有代表性的研究如日本东京大学的 Transcalibur(Shigeyama 等,2019)以及 JetController(Wang 等,2021)。Transcalibur 是一个可以手持的 2 维移动 VR 控制器,可以在 2 维平面空间改变其质量特性的硬件原型,并应用数据驱动方法获取质量特性与感知形状之间的映射关系。通过手持控制器可以有效实现用户抓握和操作物体,并且可以一定程度上降低用户的眩晕感。但手持式的触觉设备往往需要额外的定位装置,否则用户一旦在虚拟环境中放下手持式装置,便难以再次抓起。

3)机器人式触觉反馈设备。机器人式触觉反馈设备是以可移动或者可变形的机器人作为触觉代理装置,实现可移动和可变换的触觉方式。最早可以追溯到 2015 年,Cheng 等人(2015)使用 TurkDeck 的方法,借助工作人员将一系列通用模块搬运和组装为用户即将触碰到的被动实物,使用户不仅能够看到、听到,还能触摸到整个虚拟环境。Suzuki 等人(2020)在此基础之上提出了 Roomshift 方法,通过实时控制混合现实交互空间的小车来移动环境中的实物物体,提供多种交互方式。Abtahi 等人(2019b)提出了 Beyond the Force(P. Abtahi),通过可飞行的无人机作为触觉代理,提供动态的被动力触觉。图 11 所示的 4 轴飞行器目前可以支持 3 种功能:被动触觉的动态定位、纹理映射和作为可交互的被动道具。而且,无人机在交互环境中任意移动,显著地拓展了交互的空间范围。

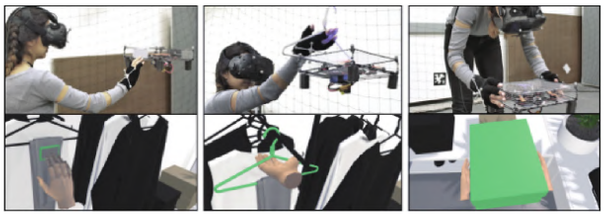


图 11　相遇型被动力触觉装置(Abtahi 等,2019b)

Fig. 11　Encounter-type haptic devices(Abtahi et al. ,2019b)

1.3.3　产业界进展

在产业界,Facebook 和 Microsoft 是研究混合现实被动力触觉交互的中坚力量。2019 年 Facebook 更新了交互装置 Tasbi,一款具有震动和挤压两种反馈方式的触觉回馈腕带。2020 年,Microsoft 提出了 PIVOT(Kovacs 等,2020),通过可变形的交互装置实现动态的相遇型触觉反馈。PIVOT 是一个戴在手腕上的触觉设备,可以根据需要将虚拟对象呈现在用户的手上。Dexmo 在 2020 年发布了新的触觉手套,Dexmo 外骨骼手套制作精良,该产品面向企业市场。Dexmo 触觉手套支持跟踪多达 11 个自由度的手势,可以灵巧地捕获用户完整的手部动作,从而使用户在虚拟环境中拥有逼真的手指感。不只是手部的力反馈,英国的 TeslaSui 生产了对全身提供被动力触觉的设备,其产品可将触觉反馈传输到身体的任何区域,从轻柔的触摸到体力的消耗感以及温度改变,并能输出运动捕捉和生物识别信息。采用带有性能监控和感觉刺激的 TeslaSuit 可以应用于公共安全、企业培训、体育和医疗康复等领域。

1.4　可穿戴交互

国际上可穿戴交互主要分为以手表手环形式为主的手势交互和触控交互的研究、皮肤电子技术与交互设计。

1.4.1　手势交互与触控交互

手势输入被认为是构成“自然人机界面”的核心内容之一,同样适用于探索可穿戴设备的输入方式。实现手势输入的关键在于传感技术,目前人机交互领域深入研究了包括基于红外光、运动传感器、电磁、电容、超声波、相机和生物信号等用于手势识别的传感技术。美国华盛顿大学和微软研究院的联合项目推进了肌肉电信号(electromyogram,EMG)在手势界面中的应用(Saponas 等,2009)。EMG 通过测量电极对之间的电势来感知肌肉活动,这可以通过侵入式(在肌肉中使用针头)或从皮肤表面进行。美国卡内基梅隆大学的 Chris Harrison 团队近些年探讨了通过在皮肤表面形成电磁场进行连续手势的识别(Zhang 等,2016;Zhou 等,2016)。通过一个戒指向佩戴的手指发出不易察觉且无害的 80 MHz、1.2 Vpp 交流信号,当用户的手指接触皮肤时,电信号会传播到手臂组织并向外辐射。信号需要时间来传播,通过测量手臂上多个电极对上的这些相位差,可以计算信号源的位置。2015 年,谷歌发布的 Soli 智能芯片运用微型雷达监测空中手势动作,可以追踪亚毫米精准度的手指高速运动(Lien 等,2016)。

系统使用高频(1～10 kHz)、150°宽的雷达脉冲,系统接收到多个动态散射中心的反射叠加,可提取移动的手的各种瞬时和动态特征,并使用机器学习技术先验捕获的训练数据集进行比较来识别手势。这种特殊设计的雷达传感器已获批被用于控制可穿戴和微型设备。

使用可伸展和贴皮式电子器件为实现皮肤界面提供了新的思路,可用于创造轻薄的电子皮肤,允许用户在其上实现触控并具有生理信号监测、视觉显示和触觉显示的功能(Withana 等,2018)。实现触觉反馈将在皮肤界面的交互里变得尤为重要,这取决于皮肤自身的触觉感知能力。德国的 Patric Baudisch 团队尝试了通过腕带手表在皮肤上实现拖动的触感,可设计简单且容易被用户感知和记忆的字符和图标(Ion 等,2015)。韩国科学技术院的人机交互团队探索了使用针阵列的触觉方式在手指上提供经过编译的信息(Je 等,2017),以及通过气流在皮肤表皮实现非接触式的压力触感(Lee 等,2016)。加拿大多伦多大学利用记忆金属在手腕上实现挤压的触觉反馈(Gupta 等,2017),通过控制驱动的线宽、力和速度产生不同感受的反馈。美国斯坦福大学的 Sean Follmer 团队通过设计手持式触觉设备来模拟虚拟操作物体的重力反馈(Choi 等,2017)。系统中两个音圈致动器通过不对称的皮肤变形产生与每个指垫相切的虚拟力,这些力可以视为虚拟物体的重力和惯性力。

1.4.2　电子皮肤交互

皮肤作为人们与外界接触的天然界面,已初步用于探索在信息交互中的作用并在若干方面的应用中体现了其优势。德国萨尔州大学的 Jürgen Steimle 团队近些年通过导电墨水、电极制作纹身纸,作为电子皮肤实现在皮肤上的显示、触摸和手势识别(Groeger 和 Steimle,2017;Olberding 等,2014;Weigel 和 Steimle,2017)。相比于触摸屏,人们在自己的皮肤上移动手指显得更加灵活,而通过纹身纸的方式使得在皮肤表面附属的设备轻而薄,更容易被用户接受。来自于该团队的一项用户研究证明,用户在皮肤上进行的触摸手势和传统触摸屏手势较为一致,但同时也因为皮肤独有的特点,用户设计出了更为丰富的触控手势,证明了皮肤作为触控界面的可行性和优势(Weigel 等,2014)。同样是对皮肤界面的探索,美国卡内基梅隆大学的 Chris Harrison 团队采取了在皮肤上投影的方式,通过肩戴投影(Harrison 等,2011)或手表微投影(Laput 等,2014;Xiao 等,2018),将手臂、手背变成显示屏,并通过深度相机或红外线等方式支持手指在皮肤表面的触控。这种方式可以更好地支持探索人们使用皮肤界面的体验,但缺点也显而易见,即需要较为复杂的投影等附属设备。同时,该团队系统地研究了把身体的各个部位当做触摸界面时的可行性和用户的喜好程度(Harrison 和 Faste,2014),对后续的研究具有参考价值。这些项目的相似之处是在皮肤上发展和拓展触控交互的模式。

另一方面,研究者也在探索皮肤界面的独特用途,比如尝试把皮肤用做设计和创作的交互平台。加拿大 Autodesk 研究院探讨了如何利用人体手臂的皮肤构建一个 3D 建模和制造的平台,并展示了以皮肤为中心的建模技术(Gannon 等,2015,2016)。韩国科学技术院的研究者们试图让用户在自己身上进行绘制来设计服装(Saakes 等,2016)。挪威代尔夫特技术大学的 Charlie C L Wang 团队则允许用户在自己皮肤和手臂上进行服装设计的同时通过热感应来分析舒适度(Zhang 等,2017c)。美国麻省理工学院(Massachusetts Institute of Technology,MIT)的 Media Lab 开展了多项以人体和皮肤为基础的概念探索项目,向人们展示了可生材料、具有生物活性材料与人体皮肤结合时产生的设计、制造以及艺术价值(Yao 等,2015)。

1.5　人机对话交互

人机对话交互过程涉及语音识别、情感识别、对话系统和语音合成等多个模块,其主要框架如图 12 所示。首先,用户输入的语音通过语音识别和情感识别模块转化为相应的文本和情感标签。而后,对话系统将其用来理解用户所表达的内容,并生成对话回复。最后,语音合成模块将对话回复转换为语音,与用户进行交互。人机对话交互的性能不仅仅

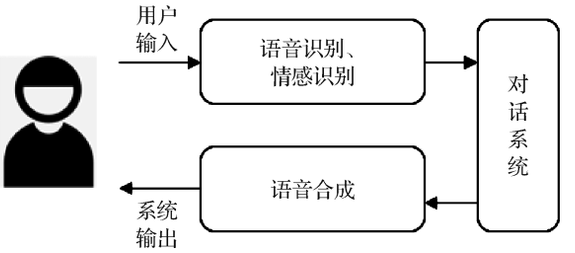


图 12　人机对话交互框架图

Fig. 12　Human-computer diaglog framework

取决于对话系统的质量,高效鲁棒的语音(情感)识别与语音合成技术对于提高用户黏性发挥着至关重要的作用。

1.5.1　语音识别

目前国际与国内对于语音识别系统的研究已经不再局限于提升识别的准确度,而是研究在更加复杂场景下的语音识别模型的表现。总体概括来看,低延迟语音识别和低资源语音识别成为研究热点。

目前国际上针对低延迟语音识别主要从两方面进行研究,一方面是研究流式语音识别,实现边听边识别,以此来降低识别出每个标记的延迟;另一方面是研究非自回归语言识别,通过摆脱解码时的时序依赖从而加速整个系统的识别速度。

针对流式语音识别的研究主要有两种思路,一种是针对 RNN-Transducer 模型进行改进,提出了表现更好的 Transformer-Transducer (Zhang 等,2020a; Yeh 等, 2019)、Conformer-Transducer (Huang 等, 2020c;Guo 等,2021)。双通解码方法(Sainath 等, 2019)的提出,进一步提升了基于 Transducer 的流式识别模型的准确率。另一种是对基于注意力机制的编码解码模型(attention-based encoder decoder, AED)的改进, 其实现思路主要是改进单调逐块注意力机制(monotonic chunk-wise attention, MoChA)(Chiu 和 Raffel,2018),其解决的主要问题是 MoChA 模型在 Transformer 上的适配以及对于通过辅助手段对模型切分编码状态的位置以及数量进行约束(Inaguma 等,2020a,b)。

针对非自回归语音识别方面的研究,国际研究上也日趋火热。非自回归语音识别因为摆脱了序列模型解码阶段的时序依赖,获得了广泛的速度提升,在自然语言处理(natural language processing,NLP)领域和语音领域均获得了很多关注。针对非自回归语音识别模型的提升整体上也是从两个角度来进行研究的,一方面是先通过编码器预测初始标签,解码器进行纠错或补全(Chi 等, 2021; Higuchi 等, 2021);另一方是通过解码器从空白序列出发,基于编码器的声学状态,预测得到完整的输出序列(Chen 等,2020a)。

近年来,国际上也掀起了针对低资源语音识别任务的研究高潮,普遍采用自监督技术或预训练技术(Schneider 等, 2019; Baevski 等, 2020a, b; Sadhu 等,2021;Hsu 等,2021)。其中最具代表性的就是 Facebook(已改名为 Meta)提出的 wav2vec 系列工作(Baevski 等,2020a,b),其将输入音频波形直接编码为声学向量表示,并通过矢量量化技术对声学向量表示进行聚类,整个预训练阶段使用对比算法进行自监督学习,然后在少量标注数据上进行微调。

1.5.2　语音情感识别

语音情感识别研究的早期阶段遵循传统的模式识别流程,即先进行特征提取,然后进行分类器设计。特征提取阶段大多依赖于手工设计的与情感相关的声学特征。大体上,这些声学特征可以分为 3 类,分别是韵律学特征、谱相关特征以及音质特征(Zhuge 等,2021)。开源工具 openSMILE(韩文静等,2014)通常用于提取一些经典的情感声学特征集。受益于深度学习革命的到来,利用深度神经网络直接从原始数据中提取特征并进行分类的端到端学习范式逐渐占据主导地位。这些研究有的从时域的原始语音信号入手(Eyben 等,2010),有的则从频域的语谱图入手(Tzirakis 等,2018),此外也有一些研究同时结合两者进行端到端的语音情感识别(Li 等,2018)。由于语音情感识别的数据库通常都比较小,人工设计的深度神经网络往往容易过拟合,因此学习到的声学情感表征可能会面临着泛化能力不足的问题。为此,一些研究(Hershey 等,2017;Zhao 等,2018)采用在大规模音频数据库上预训练的深度神经网络(如基于卷积神经网络的 VGGish(Bakhshi 等, 2020)、Wavegram-Logmel-CNN (Zhao 等, 2018)和 PLSA(pretraining, sampling, labeling, and aggregation)(Kong 等,2020),以及基于 Transformer 的 AST(audio spectrogram Transformer)(Gong 等, 2021b)等)进行特征提取,当然也可以继续在语音情感识别数据库上进行微调。受益于最近大规模无监督预训练的兴起,目前已有不少研究采用自监督学习的方式从大量未标注的语音数据中提取有用的音频表征并用于下游的情感识别任务,如 Mockingjay(Liu 等,2020),Tera(Liu 等,2020),wav2vec 2.0 (Liu 等,2021;Baevski 等,2020b;Pepino 等,2021)等。此外,为了挖掘语音信号中的语义信息,也有一些研究同时结合声学信息和文本信息进行多模态语音情感识别的研究(Li 等, 2021; Yoon 等, 2019, 2020)。

1.5.3　语音合成

目前语音合成研究主要集中在韵律建模、声学

模型以及声码器等模型的建模之中,以提高合成语音的音质和稳定性,并提高在小样本数据集上的泛化性。具体地,谷歌 Deepmind 研究团队提出了基于深度学习的 WavetNet(van den Oord 等,2016)语音生成模型。该模型可以直接对原始语音数据进行建模,避免了声码器对语音进行参数化时导致的音质损失,在语音合成和语音生成任务中效果非常好。2017 年 1 月,Sotelo 等人(2017)提出了一种端到端的用于语音合成的模型 Char2 Wav,其有两个组成部分:一个读取器和一个神经声码器。读取器用于构建文本(音素)到声码器声学特征之间的映射;神经声码器则根据声码器声学特征生成原始的声波样本。本质上讲,Char2 Wav (Sotelo 等,2017)是真正意义上的端到端语音合成系统。谷歌科学家提出了一种新的端到端语音合成系统 Tacotron(Wang 等,2017b),该模型可接收字符的输入,输出相应的原始频谱图,然后将其提供给 Griffin-Lim 重建算法直接生成语音。此外,由于 Tacotron 是在帧层面上生成语音,所以它比样本级自回归方式快得多。研究人员进一步将 Tacotron 和 WaveNet 进行结合(Shen 等,2018),在某些数据集上能够达到媲美人类说话的水平。为了提高合成效率,一些声码器加速工作也有显著进展(Valin 和 Skoglund,2019;Yamamoto 等,2020)。

1.5.4 对话系统

对话系统从应用角度划分,可以分为任务型对话系统和闲聊型对话系统;从方法上划分,可以分为基于管道的方法和基于端到端的方法。基于管道的方法需要分别实现自然语言理解、对话管理和自然语言生成 3 个模块,最终形成一个完整的系统。这种模块级联的方式会导致误差传递问题,因此基于端到端的方法目前成为主流的对话系统方案。

为克服端到端对话系统中存在知识难以融入学习框架的问题,Eric 等人(2017)引入键值检索网络整合知识库信息。Madotto 等人(2018)提出了 Mem2Seq(memory to sequence)模型,采用指针网络实现将知识库嵌入到对话系统中。Wu 等人(2019)改进了 Mem2Seq 模型,提出 GLMP(global-to-local memory pointer)模型,将外部知识融入对话系统之前进行过滤,并且加入了骨架循环神经网络机制生成对话模板。

除了基于文本的对话系统,学者们在多模态对话系统方面做了许多工作。Barbieri 等人(2018)根据对话上下文预测 emoji 表情。Haber 等人(2019)设计了一种对话系统,让用户使用自然语言与机器谈论给定的视觉内容。

1.6 多模态融合

如何将不同模态的信息在人机交互系统中有效融合,提升人机交互的质量,同样值得关注。多模态融合的方法可分为 3 种:特征层融合方法、决策层融合方法以及混合融合方法(Debie 等,2021)。3 种融合方法如图 13 所示。特征层融合方法将从多个模态中抽取的特征通过某种变换映射为一个特征向量,而后送入分类模型中,获得最终决策;决策层融合方法将不同模态信息获得的决策合并来获得最终决策;混合融合方法同时采用特征层融合方法和决策层融合方法,例如可以将两种模态特征通过特征层融合获得的决策与第 3 种模态特征获得的决策进行决策层融合来得到最终决策。

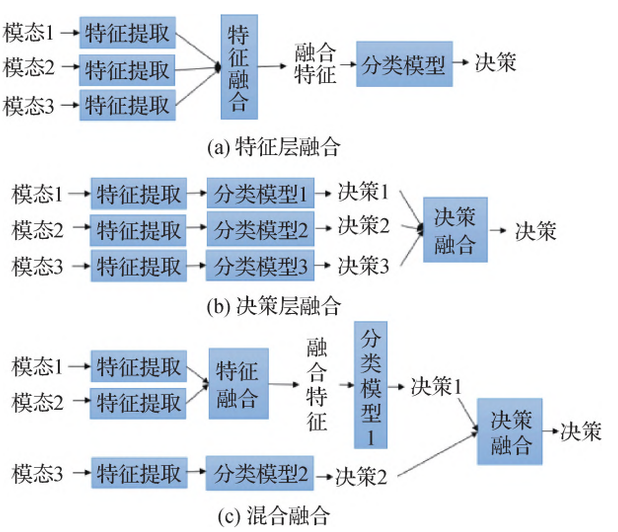


图 13 3 种不同的多模态融合方法

Fig. 13 Three different multi-modal fusion methods((a) feature level fusion;(b) decision level fusion;(c) hybrid fusion)

2 国内研究进展

2.1 大数据可视化交互

2.1.1 大数据可视化设计

在大数据可视化领域,国内的发展也已经逐渐走向成熟,每年都有许多可视分析系统不断涌现(Chen 等,2021;Wang 等,2021;Deng 等,2021)。近年,沉浸式大数据可视化得到了发展,浙江大学的

Ye 等人(2021)探索了如图 14 所示的无缝结合羽毛球比赛数据绘制的 2D 和 3D 可视化视图的问题，Chu 等人(2022)探索了结合高度来凸显羽毛球数据中多个战术之间存在的差异性问题，如图 15 所示。由此可以看出，沉浸式大数据可视化对数据分析和展示问题提出了有效的解决方法。

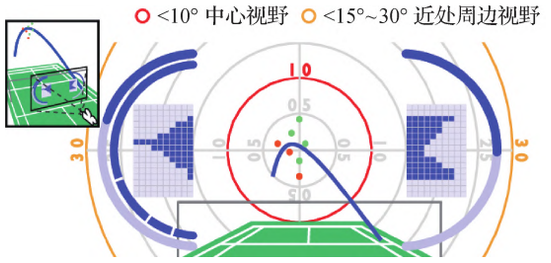


图 14　2D 和 3D 可视化结合的设计(Ye 等,2021)

Fig. 14　Design combining 2D and 3D visualization(Ye et al. ,2021)

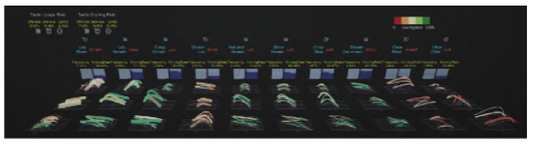


图 15　3 维羽毛球战术可视化(Chu 等,2022)

Fig. 15　3D badminton tactic visualization(Chu et al. ,2022)

但是相比国外学者在沉浸式大数据可视化上的研究，国内仍处于起步阶段，所以接下来的发展还需要各高校继续深入研究。

2.1.2　非视觉感知的交互辅助

大数据可视化交互中，非视觉的感知交互方式以触觉最为常见。通过反馈力的大小与方向，用户可以使用触觉直观地感知到连续的高维数据信息。赵俭辉等人(2021)使用电磁力反馈设计了一种交互方法，并解决了虚拟手术中沉浸感不足的问题。如图 16 所示，用户在虚拟手术中操作的导丝可以获得真实手术环境下的多种反馈力，同时用户也获得了更逼真的手术体验，提高了术前虚拟训练的效果。在一部分沉浸式系统中，用户的 3 维感知也在分析中发挥了重要的作用。如图 17 所示，杭州师范大学的潘志庚等人(2021)通过一种数字对象和真实物体的孪生配准技术将虚拟世界中的物体渲染到真实世界中，从而将多种分析对象放置于分析者身旁。该技术充分利用分析者对于分析对象的 3 维感知。在教学实验中，该技术可以辅助参与者有效地进行

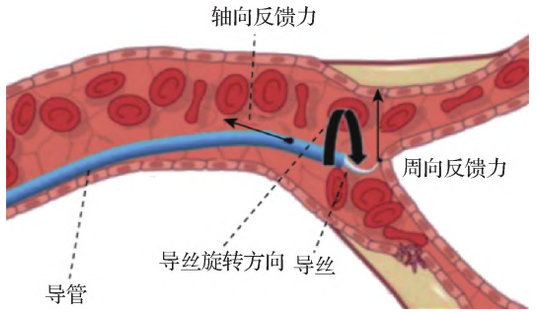


图 16　介入手术中导丝导管力反馈(赵俭辉 等,2021)

Fig. 16　Force feedback of guidewire during interventional surgery(Zhao et al. ,2021)

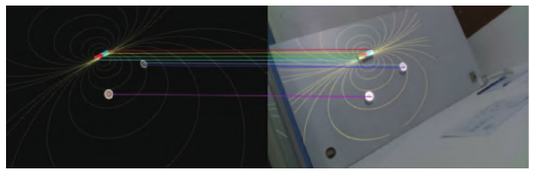


图 17　数字对象(左)以及渲染出的真实对象(右)(潘志庚 等,2021)

Fig. 17　Digital object (left) and the rendered real object (right)(Pan et al. ,2021)

磁感线的交互式学习。近几年，国内围绕嗅觉、听觉等通道的成果还较为匮乏并落后于国外。这些感知通道还需要研究者进一步探索其交互方式以及配套硬件设施。

1)基于接触的交互。触控式大屏的出现对多人协同分析大数据可视化起到了促进的功能。仁光科技先后设计了 13 种自然交互对触控式的可视化大屏进行数据操作，例如手指触控、笔触触控，通过触控式交互可完成对数据的选择、可视化的拖拽缩放等。

2)基于手势的交互。西南科技大学的 Wang 等人(2020a)提出了一套基于手势的“所见即所得”的交互方式，可完成对体数据进行抓取移动等动作，与在真实场景里的交互动作一样。浙江大学的 Ye 等人(2021)设计了具象化的羽毛球挥拍隐喻用于飞行轨迹的筛选，羽毛球分析专家挥动 VR 手柄，根据空气动力学，系统会基于手柄的移动方向和移动速度模拟一条虚拟的羽毛球轨迹，并从已有数据中查询到与之相似的轨迹并进行分析。

3)基于注视的交互。视线追踪技术捕捉人们的视线焦点，可以代替手部对数据进行选择的操作行为，减少手部交互带来的疲劳。Hu 等人(2021)提出的 FixationNet 可以根据人们的历史凝视位置、

探索任务的对象以及用户的头部移动速度,预测其在 VR 中近期的注视情况,FixationNet 提高了任务为导向的可视分析系统中用户的探索能力。

4)基于移动导航的交互。移动是沉浸式大数据可视化中最常用的交互之一,山东大学的 Li 等人(2020)提出了一种重定向行走方法,支持用户在物理空间中行走较小的距离,同时在虚拟环境里完成远距离的行走,解决了物理空间有限的问题。他们提出了一种基于 Voronoi 的方法来生成行走路径,并且采用重定位和曲率调整的静态图映射方法将虚拟空间的行走路径与物理空间进行映射,由此实现在物理空间中的连续移动,拓展了人们在沉浸式环境中的探索空间。

2.2 基于声场感知的交互

2.2.1 基于声场感知的动作识别

国内对于声场识别手势的研究相对较少。其中,国内学者提出的 PrivateTalk(Yan 等,2019)利用双耳耳机上的麦克风识别出了用户捂嘴时的语音,实现了语音交互唤醒的优化。魏文钊和何清波(2018)设计出了一套基于超声波识别手势交互的系统。

2.2.2 基于声源定位的交互技术

ReflecTrack(Zhuang 等,2021)利用工作生活中的反射面,使用智能手机上的双麦克风实现了 22.1 mm 精度的 3 维声学定位。该工作使用频率超出人耳听觉范围的 FMCW 声音信号,同时识别直接路径的声音信号和反射得到的声音信号,实现了只需要两个麦克风的声学定位技术。由于反射面在生活中很常见,基于该技术可以实现运动追踪和精细的手势识别等多种交互技术。FaceOri(Wang 等,2022b)利用任意智能设备的扬声器发出频率超出人耳听觉范围的 FMCW 声音信号,通过使用用户双耳佩戴的主动降噪耳机上的麦克风,利用超声波测距方法,创新头部相对智能设备的头部空间位置与角度的精准连续追踪技术,支持包括交互对象感知与健身动作识别等更加智能高效的人机交互技术。

2.2.3 基于副语音信息的语音交互增强

Qin 等人(2021)提出了基于单麦克风近距离风噪特征的凑近免唤醒语音交互技术 ProxiMic,可用于手机手表耳机的手持或穿戴设备的凑近即说,该工作利用人距离麦克风近距离状态下说话的自然吐气特征,设计了基于风噪一致性的两步算法,用户轻声或气声亦可激活系统,具有私密性强、鲁棒性高和准确率高等特点。

2.2.4 普适设备上的音频感知与识别

国内在智能手机上的音频感知与识别研究较多,典型的如李凡等人(2021a,b)提出的两种在驾驶环境下进行音频感知与识别的工作:1)利用智能手机扬声器收集并基于自适应子带谱熵方法和神经网络进行驾驶环境下的呼吸道症状检测技术(李凡等,2021a);2)利用智能手机扬声器与麦克风组成的声呐系统,基于物理原理实现车辆行驶速度的检测方法(李凡 等,2021b)。此外,陈超(2021)提出一种利用智能手机内置扬声器与麦克风实现对疲劳驾驶行为感知的检测技术。

2.3 混合现实实物交互

国内在被动力混合现实交互方面跟进较为迅速,与国际上的前沿水平相差不大。目前,虚拟现实用户主要通过视觉、听觉感知环境,而具有双向信息传递能力的触觉通道往往得不到支持。触觉呈现的功能缺失或位置精准度低下会造成用户对虚拟物体的感知失真,降低用户使用沉浸感。同时,触觉引导的欠缺也会导致用户交互效率大幅降低。北京理工大学、北京航空航天大学、中国科学院软件研究所和中国科学院大学等项目组,针对长时虚拟沉浸中的显触失配问题,提出了一系列基于被动触觉的虚实融合技术。

2.3.1 静态的被动力触觉

在静态的被动力触觉方面,Zhao 等人(2021)提出基于纹理感知特征的触觉信号采集方法,基于静电振动触觉显示技术及最小可觉差估计法的触觉感知信号量化、采集多通道纹理信息,并在此基础之上进一步提出基于静电力反馈的触觉渲染方法,提高虚拟纹理真实感。Guo 等人(2020)提出基于实例分割的被动触觉对象选择性渲染及特征化表达方法,平衡虚实融合系统中环境沉浸感与系统交互效能冲突,扩展虚实融合场景渲染自由度,实现虚实空间智能化融合。

2.3.2 相遇型被动力触觉及 3 种触觉设备

在相遇型被动力触觉方面,如图 18 所示,Jiang 等人(2019b)提出了 HiFinger 方法。HiFinger 是一种单手可穿戴的文本输入技术,可通过拇指对手指的触摸实现输入时的触觉反馈以及快速、准确、舒适地输入文本,适用于用户需要在虚拟环境中移动(如行走)的移动场景,有效地提供了一种混合现实

环境中的输入解决方案。Zhang 等人(2019)也开发了一种轻量的多指力反馈手套,通过一种在每个手指关节上使用分层干扰片的解决方案,在增强虚拟现实和远程操作系统的保真度方面具有巨大应用潜力。Li 等人(2020)针对难以在虚拟环境中添加真实物体的问题,提出了一种基于连杆机构的原型框架 HapLinkage。该框架提供了典型的运动模板和触觉渲染器,便于虚拟手动工具的代理设计。机械结构可以很容易地修改,能够轻松快速地为各种混合现实场景创建手动工具的原型,并赋予它们动力学和触觉特性。Xue 等人(2019)提出了 MMRPet(modular mixed reality pet),一种可通过磁力组装的虚拟宠物交互装置,模拟逼真的被动力触觉。通过将虚拟宠物叠加在被跟踪的宠物实物上,兼具丰富的视觉信息和实物交互,同时宠物实物采用模块化的结构设计,各模块能够以不同方式相连接,构成不同形态结构的宠物实物,避免不同的虚拟宠物均需要一个单独的宠物实物作为被动力触觉反馈的提供者,使被动力触觉反馈方案更加灵活,同时赋予用户更多的交互自由。

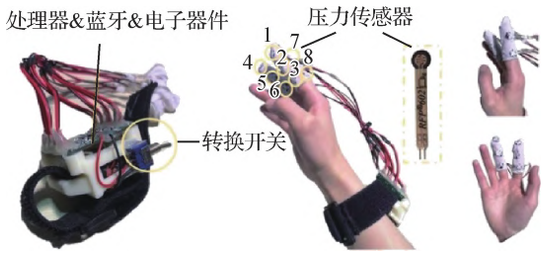


图 18　力反馈输入装置(Jiang 等,2019b)

Fig. 18　Force feedback input device(Jiang et al. ,2019b)

2.3.3　产业界进展

在产业界,国内目前的发展较为迟缓,目前还没有非常完善的混合现实触觉解决方案。国内主要提供混合现实设备的公司,目前在触觉外设方面仍处于探索阶段。除 HTC Vive 的控制手柄之外,研究人员曾经提出过一种沉浸式地板。地板上安置有大量电动微型模块,它们会根据用户在混合现实中的内容改变地板的表面形状,提供一定的被动力反馈。此外,PPGun VR 曾推出过一款枪型控制器,便于优化用户在虚拟环境中的射击体验。通过与主机相连的仿真步枪,用户可以真实地完成射击、填装子弹等一系列操作。但由于触觉代理对象种类繁多而且形状复杂,目前混合现实中的触觉并没有一套产业化的解决方案。

2.4　可穿戴交互

国内对可穿戴设备交互的研究主要集中在新型的传感技术来支撑手势、语音等交互行为,以及对交互意图理解和交互界面的优化等方向。中国科学院计算技术研究所陈益强团队从事普适计算的研究,包括用手表内置传感器进行用户手臂动作的捕捉,并依次进行用户动作建模及拓展其在空间环境里的交互场景(Wang 等,2019;Chen 等,2020b)。北京大学张大庆团队利用可穿戴设备和 WiFi 信号解析,对用户在空间中的动作和其自身的生理指标进行监测(Yang 等,2015;Wang 等,2016)。如提出一种基于转换的分割方法,利用一对接收器天线上的相位差方差作为显着特征,自动分割连续捕获的 WiFi 无线信号流中的所有跌倒和类似跌倒的活动。南京大学谢磊团队等对以可穿戴 RFID(radio frequency identification)标签为基础的无线信号传感进行建模和解析,支持用户动作和行为的检测(Xie 等,2010;Wang 等,2018)。系统中只在标签阵列后面部署一根 RFID 天线,持续测量标签阵列发出的信号,根据相应的信号变化识别手势,并将多根手指作为一个整体进行识别,然后提取多根手指的反射特征作为图像。

同时,国内的研究重视对人因元素的考虑和对用户行为的建模。清华大学史元春团队研究手表等小型触摸屏上的文字输入问题,通过新型的表盘界面设计与用户意图推理等技术的结合,创造出高效的文字输入技术(Yi 等,2017;Han 等,2018)。要输入文本,用户可以转动表圈,用光标敲击圆形键盘上的键,为了最小化旋转距离,每个光标的位置在每次按键选择后根据需要下一个按键的概率进行动态优化。中国科学院软件研究所田丰团队在设备周围的手势交互技术实现和高效的适用于小屏幕的手表命令界面的设计等方面进行了创新(Han 等,2017,2018)。如一种新的手势是通过将屏幕上的一个角拖动到不同的方向和距离来执行的,每个角都映射到某个命令,并且可以卷曲/剥离以浏览命令下可用的值。Robin Bing-Yu Chen 团队研究了用手掌和指间作为触摸界面在手势输入和文字输入等方面的应用(Huang 等,2016;Wang 等,2015)。该工作解决了两个人体工程学因素:手部解剖结构和触摸精度。手部解剖结构限制了拇指的可能运动,这进一步影

响了交互过程中的身体舒适度。触摸精度是一个人为因素,它决定了用户可以如何精确地操作设置在手指上的触摸小部件,以及小部件的有效布局。清华大学史元春团队同时在触控、手势和语音等多模态输入通道下交互行为优化和自然等方面做出创新(Qin 等,2021)。如用户可以将嵌入麦克风的设备放在嘴边,并直接对着设备说话,而无需使用唤醒词或按下按钮,为了检测靠近麦克风的语音,系统使用了用户说话并向麦克风吹气时观察到的爆音的特征。

2.5 人机对话交互

2.5.1 语音识别

国内与国外针对语音识别的研究整体趋势是趋同的,但是在聚焦的技术方面还是存在一定的差异。国内的实验研究也紧跟低延迟语音识别和低资源语音识别两个方向。

针对低延迟语音识别方面,国内以中科院、清华大学和西北工业大学为代表,围绕非自回归语音识别模型做了不少探索性的工作;在流式语音识别方面,国内划分成3种思路:1)字节跳动公司、腾讯公司和中国科学院自动化研究所对 Transducer 模型进行了实用化的改进,提升识别速度和准确率(Huang 等,2020b;Tian 等,2019,2021b;Tian 等,2021a);2)百度公司聚焦于使用 CTC(connectionist temporal classification)模型对连续编码状态进行切分,然后使用注意力模型进行解码,先后提出了 SMLTA(streaming multi-layer truncated attention model)和 SMLTA2(http://research.baidu.com/Blog/index-view?id=109);3)中国科学院自动化研究所、出门问问公司和阿里巴巴公司尝试实现将流式模型和非流式模型统一到一个框架中(Tian 等,2020;Zhang 等,2020b)。

受限于计算资源和数据规模,国内高校科研单位对于自监督与无监督语音识别的研究偏少,这部分研究主要集中于企业,有京东公司、字节跳动公司、猿辅导和滴滴,其工作偏向跟随和扩展性质,其代表工作包括 SCALA(supervised contrastive learning)和 BERT(bidirectional encoder representations from Transformers)的变体(Jiang 等,2019a,2021;Fu 等,2021)。

2.5.2 语音情感识别

国内语音情感识别的研究早期阶段也集中在区分性语音情感特征的提取以及分类器的设计(Sun 等,2021;赵力 等,2004;金学成,2007)。如,东南大学的赵力团队(Sun 等,2021)在2004年提出了一种利用全局和时序结构的组合特征以及 MMD(modified Mahalanobis distance discriminant)进行语音情感特征识别的方法。而后受益于深度学习的发展,一些新型的深度神经网络逐渐用于语音情感识别并取得了不错的效果,包括深度信念网络(韩文静 等,2008)、基于高效通道注意力的 CRNN(convolutional recurrent neural network)(韩文静 等,2014)和 Sinc-Transformer(戴研研 等,2021)等。在数据库建设方面,中国科学院自动化研究所录制了 CASIA(Institute of Automation, Chinese Academy of Sciences)汉语情感语料库,该数据库涵盖了4位录音人在纯净录音环境下以5类不同情感演绎的9 600句语音。

2.5.3 语音合成

在语音合成领域,国内研究与国际基本保持一致。为了提高模型的鲁棒性,百度公司提出了 Deep Voice 和支持多说话人的 Deep Voice 2(Arık 等,2017),它通过相应的神经网络代替传统参数语音合成流程中的每一个组件。为了提高模型在小数据上的泛化性(Jia 等,2018),中国科学院自动化研究所等科研机构通过将目标说话人的韵律与音色进行解耦(Wang 等,2020b),提高模型的泛化性,在小数据集的目标说话人上表现良好。国内各大互联网厂商也陆续推出基于个性化语音合成的算法服务,有助于语音合成推广到更加广泛的领域。

2.5.4 对话系统

针对融合知识的端到端对话系统,哈尔滨工业大学的研究人员改进了 Mem2Seq(Madotto 等,2018)模型中存在的实物生成不一致的问题(Qin 等,2019),并且提出动态融合网络(Qin 等,2020)以提高对话系统的集外迁移能力。中国科学院自动化研究所的研究人员(Chen 等,2019)提出采用一种心理学模型将外部知识与端到端对话模型进行有机融合。

针对多模态对话系统,山东大学的研究人员(Cui 等,2019)提出 UMD(user attention guided multimodal dialog system)模型,利用多模态编码器和解码器分别编码多模态话语和生成多模态响应。中国科学院计算技术研究所的研究人员(Debie 等,2021)建立一种开放域多模态对话数据集,推动了

多模态对话系统的发展。

3 国内外研究进展比较

3.1 大数据可视化交互

在大数据可视化交互中，由于是在传统可视化的基础上发展起来，面向可视化设计、交互设计的研究比较早，研究成果相对成熟，但受限于平面化的展示空间与交互空间。沉浸式技术的发展拓展了数据的呈现，支持数据的 3 维可视化，由此释放了人们的立体视觉。受益于硬件技术的进步，非视觉的交互技术陆续提出并用于辅助视觉交互。头戴式设备、触摸式大屏和传感器的发展，为多模态可视化交互创造了可能性。国内外研究进展对比见表 1。

表 1 大数据可视化交互国内外研究进展对比

Table 1 Comparison of domestic and foreign research progress on big data visualization interaction

国外代表性研究工作		国内代表性研究工作		国内外对比
研究机构	工作内容	研究机构	工作内容	国内外在大数据可视化交互的研究上较成熟，国内基于不同感知通道的交互设计研究较为滞后，其中围绕听觉、嗅觉等通道的成果尤为匮乏。
澳大利亚莫纳什大学 巴西南大河联邦大学	沉浸式地图可视化交互 时空轨迹的探索	浙江大学	传统的平面式可视分析 沉浸式交互	

最早沉浸式可视化是在 IEEE VIS（Visualization conference）2014 年的研讨会上提出，自此，国外有大量学者开始探索沉浸式可视化，并提出了系列的可视化设计、基于不同感知通道的交互设计（Siu 等，2020；Prouzeau 等，2019；Patnaik 等，2019）以及多模态融合的交互设计，例如澳大利亚莫纳什大学的 Tim Dwyer 教授带领的团队对沉浸式地图可视化交互提出了系列工作，包括利用高度编码起讫点流图（Yang 等，2019）、对地图视图进行操作的基于接触的交互设计（Yang 等，2021b）等，巴西南大河联邦大学的 Jorge Wagner 为时空轨迹可视化的探索提出了虚拟桌面的隐喻（Filho 等，2019）、对轨迹移动旋转筛选的一整套手势交互设计（Wagner 等，2021）。

然而，国内的大数据可视化交互研究主要集中在传统的平面式可视分析中，沉浸式交互集中在工业场景应用中，沉浸式可视化的发展带动了国内学者对可视化交互的研究。浙江大学巫英才团队开展了羽毛球轨迹分析（Ye 等，2021）与战术分析（Chu 等，2022）的工作，将羽毛球轨迹还原在 3 维空间中，利用挥拍隐喻对轨迹进行筛选，通过小倍数图布局对包含时序信息的战术数据进行展现，提高了专家数据探索的能力。相较于国外，国内基于不同感知通道的交互设计研究较为滞后，其中围绕听觉、嗅觉等通道的成果尤为匮乏。

总体而言，国内外在传统的大数据可视化交互的研究上较成熟，在基于多模态交互的沉浸式可视化的研究中，国外学者的工作较为多样，国内学者的研究较为单一，但是考虑到沉浸式可视化仍处于发展阶段，国内和国外的学者应该相互学习，推动该领域共同发展。

3.2 基于声场感知的交互

国内外研究进展见表 2。

在基于声场感知的动作识别方面，国内的研究相对较少，而国外对于利用耳机、腕带和手表等可穿戴设备进行动作识别有更加充分的探索。应用麦克风和陀螺仪等传感器，既实现了精细的手势交互动作，也对用户与其他物品的交互方式进行了研究。关于基于声源定位的交互技术，国内近期的相关工作减少了定位所需的麦克风数量，使日常场景下基于耳机和智能手机的定位成为可能。国外研究则对利用不同声学测距方法实现交互技术进行了更广泛的探讨，在被追踪设备主动和被动发声两个方向都进行了探索。在基于副语音信息的语音交互增强方向，国内近期工作实现了智能设备的凑近即说；国外研究则以多种方式利用语言中的非语言信息，加强了人机间的语音互动。对于普适设备上的音频感知与识别技术，国内工作主要集中于利用智能手机上的扬声器和麦克风来进行识别，但基于其他设备的研究较少；国外工作在利用手机、腕部设备等实现生理感知和环境识别等方面都有涉及。整体而言，国内在基于声场感知的交互技术方面虽然近些年发展较快，但是整体在技术深度与应用广度上仍然落后于国际先进水平。

表 2　基于声场感知的交互国内外研究进展对比

Table 2　Comparison of domestic and foreign research progress on interaction based on sound field

	国外主要研究工作	国内主要研究工作	国内外研究对比
基于声场感知的动作识别	应用麦克风器和陀螺仪等传感器,既实现了精细的手势交互动作,也对用户与其他物品的交互方式进行了研究。	相对较少。	国内在基于声场感知的交互技术方面虽然近些年发展较快,但是整体在技术深度与应用广度上仍然落后与国际先进水平。
声源定位交互	聚焦声学测距方法实现交互技术。在被追踪设备主动和被动发声两个方向都进行了探索。	减少了定位所需的麦克风数量,使基于耳机和智能手机的定位成为可能。	
副语音信息的语音交互增强	以多种方式利用语言中的非语言信息,加强人机间的语音互动。	智能设备的凑近即说。	
音频感知与识别	利用手机、腕部设备等实现生理感知和环境识别等方面都有涉及。	利用智能手机上的扬声器和麦克风来进行识别,但基于其他设备的研究较少。	

3.3　混合现实实物交互

国内外进展见表 3。总体而言,国内外在混合现实中的被动力触觉方向,研究进展较为类似,但研究重点略有不同。在科学研究中,国际上在相遇型触觉方面,通过使用一个或多个机器人协同控制,实现动态模拟交互空间的变化方面有着明显的优势。国外的机器人产业比较发达,可用的触觉代理往往多种多样,比如各种大小的机器人、小车和无人机等。此外,由于更高精度的定位设备的研究比较成熟,国际上对于大范围空间的交互进行了更多研究,而国内的研究往往是在用户面前的较小范围。国内的研究更关注于交互装置,以及如何通过单一、简单的交互装置来实现多种形式的触觉。通过少量或者简单的触觉代理,实现更为复杂的功能。除此之外,国内已经完成了相当一部分交互的测试工作,具有一定的参考价值。在产业界,国际上已经有了手部可穿戴式、全身可穿戴式和腕部携带式 3 种比较主流的产品,可提供力反馈甚至热反馈,有丰富的触觉交互内容。然而国内产业几乎没有混合现实触觉的解决方案,因此在触觉方面的设备比较少。伴随着虚拟现实中触觉技术的不断发展,相信国内产业界将在此方面有一定进展。

表 3　混合现实实物交互国内外进展比较

Table 3　Comparison of domestic and foreign research progress on mixed reality physical interaction

	国外主要工作	国内主要工作
学术	1)通过使用一个或多个机器人协同控制,实现动态模拟交互空间的变化方面有着明显的优势; 2)可用的触觉代理形式多样,比如各种大小的机器人、小车和无人机等; 3)由于更高精度的定位设备的研究比较成熟,国际上对于大范围空间的交互进行了更多的研究。	1)更关注于交互装置,以及如何通过单一、简单的交互装置来实现多种形式的触觉。通过少量或者简单的触觉代理,实现更为复杂的功能; 2)国内已经完成了相当一部分交互的测试工作,具有一定的参考价值。
产业	已经有了手部可穿戴式、全身可穿戴式和腕部携带式 3 种比较主流的产品,可提供力反馈甚至热反馈,有丰富的触觉交互内容。	几乎没有混合现实触觉的解决方案,因此在触觉方面的设备比较少。

3.4　可穿戴交互

国内外研究进展见表 4。

3.5　人机对话交互

国内外研究进展见表 5。

4　结　语

本文系统综述了多模态人机交互的发展现状和

表 4　可穿戴交互国内外研究现状

Table 4　Comparison of domestic and foreign research progress on wearable interaction

	国内外研究对比
国外的优势	1)国外提出了具有前瞻性的一些设计概念,并通过原型进行了验证,如通过电磁信号、生理电信号和柔性电子技术等构建的偏向于未来的交互技术探索,对思考可穿戴交互的新形式和可能性具有较大的引领作用; 2)国内在这一方面的研究较少;更多的研究重点在基于已有的内置传感器和信号源,通过算法的方案进行信号的解析和优化,从而对交互行为进行推理。
国内的优势	国内在用户行为建模和用户意图理解,以及基于此的交互界面的创新设计、多模态自然人机交互等领域的研究处于领先位置。
国内外需进一步提高的方面	对皮肤界面的探索还处于初级阶段。其具有广阔的探索空间和许多需要考虑的关键问题,比如如何发挥皮肤的特有属性拓展其应用场景,探索皮肤界面的新的实现形式,规划统一的用户体验以及皮肤界面的可接受度、可用性和安全性等。

表 5　人机对话交互国内外研究进展对比

Table 5　Comparison of domestic and foreign research progress on human-machine dialogue interaction

	国内外研究对比	
	国内	国外
语音识别	1)研究更加偏向实用性质; 2)以中国科学院自动化研究所、清华大学、西北工业大学为代表,受限于数据量和计算资源,主要研究聚焦低延迟语音识别与复杂场景下的识别;企业界则以腾讯、字节跳动、京东等公司作为代表,更加关注通用语音识别性能提升,也进行一部分自监督、无监督工作的预研。	1)研究更加偏向研究性质; 2)研究机构以 Google 和 Facebook 为代表,聚焦在流式语音识别研究和自监督无监督语音识别研究方向。
语音合成	国内外研究进展同步,在合成语音的主观评分、模型训练效率以及在小样本数据集上的表现力均能达到前沿水平。	
对话系统	国内外的研究处于并跑阶段。	
情感识别	国内语音情感识别的研究基本上延续了国际上相关研究的发展路线。相较于国外,目前国内关于语音情感识别的发展还有待进一步提升。	

新兴方向,深入梳理了大数据可视化交互、基于声场感知的交互、混合现实实物交互、可穿戴交互和人机对话交互的研究进展和国内外研究进展比较。针对每项研究内容的发展趋势与展望如下:

4.1　大数据可视化交互

在大数据可视化交互中,可视化设计的研究发展较早,成果比较成熟,然而如何利用人们的多感知通道提出交互设计,以增加对数据可视化的理解促进研究,是目前的研究热点之一。触觉、听觉等感知辅助可以减轻数据遮挡带来的观察不便,但是这又可能带来用户移动交互上产生的空间范围小、易发生碰撞等问题。因此,各模态的交互组合、适用的分析任务以及效率问题仍有待探索。

另外,由于目前设备的固有限制,人们在做出交互行为时,低精度的识别算法会影响分析效率,同时当人们长时间佩戴头戴式设备时,会出现疲惫与不适感。识别算法的提高、无形的交互动作和有形用户界面的合理结合以及设计可以减轻用户疲劳的手势组合,也是未来需要攻克的问题。

4.2　基于声场感知的交互

智能手机、手表和耳机等普适设备持有量持续快速增长,利用这些设备进行声场感知来提升用户的交互体验将成为一种趋势。现有工作主要面向单一设备开展研究,对跨设备的联合感知研究相对匮乏。然而,跨设备感知可以有效地扩展感知通道,实现对交互意图在感知能力上的提升,因此,基于跨设备分布式声场感知的交互技术将会是一个新的发展趋势。此外,类似智能耳机、智能音箱等设备的大规模使用,空间中麦克风具有常开特性,如何实现隐私保留的全域感知(全屋感知等)将成为另一个发展方向。利用房间中的声音信号,既可以实现实时的手势识别、运动追踪,也可以对人的生理信号、健康

状况进行监测。使多种设备连结起来共同感知人和环境、实现跨设备的交互技术,将减少交互路径、使交互体验更加自然高效。

4.3 混合现实实物交互

基于被动力触觉的混合现实交互,就交互对象而言,是从单一的静态交互物体,逐渐向多个物体、多样化物体、可移动的交互对象、可变形的交互装置以及可提供动态力反馈的方向发展。

受益于科技的发展,多模态同步混合现实很有可能发展为混合现实中人机交互的主要模式。多模态同步混合现实是虚拟世界与现实世界相结合的统一概念,为理解和设计连接虚拟世界和现实世界的各种系统提供了一些思路。系统将被动力触觉和主动力触觉相结合,可以给用户更好的交互体验。交互的触觉代理会更小型化、更易获得、甚至就是日常生活中常用的物品。综上所述,触觉反馈在混合现实中有着重要的地位,并在未来有着很大的应用前景。

4.4 可穿戴交互

智能穿戴设备正逐步成为普适计算的载体和方式之一,朝着微型化、集成化、依赖无所不在的实时网络和传感器获取数据、通过大量数据的实时采集和计算分析、通过增强的视觉和触觉感官及认知体验来实现设备与用户、设备与环境、以及用户与环境之间的自然交互发展。面对智能穿戴技术迅猛发展和用户需求增加,必须提升已有的智能穿戴人机交互技术,拓展新的交互通道和交互方式,拓宽人机数据沟通渠道,增强设备采集和处理生物信号能力,探索高效自然的关键交互原则和交互技术。

4.5 人机对话交互

语音识别方面,自回归语音识别模型能够极大地降低系统的延迟,在非流式识别场景具有重要的应用价值,但是性能还有待提升;噪声、多说话人和说话人重合等复杂场景下的语音识别准确率需要进一步提高。语音合成方面,现有语音合成技术主要存在两方面的挑战:一是自然口语声音的伪造很难接近真人;二是资源受限条件下伪造声音的自然度和可懂度下降明显。进一步提高自然口语声音的合成自然度和提升资源受限条件下合成声音的音质是语音合成的未来发展趋势。在语音情感识别方面,学习范式上从监督学习逐渐过渡到基于大规模无标注数据进行预训练的无监督学习。对话系统方面,多模态预训练模型(Fei 等,2021)蓬勃发展,将多模态预训练模型的强大表征能力与对话系统结合,来提高多模态对话系统的性能将是未来值得探索的方向。

致　谢　本文由中国图象图形学学会人机交互专业委员会组织撰写,该专委会更多详情请见链接:http://www.csig.org.cn/detail/2490。

参考文献(References)

Abtahi P, Gonzalez-Franco M, Ofek E and Steed A. 2019a. I'm a giant: walking in large virtual environments at high speed gains//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #522 [DOI: 10.1145/3290605.3300752]

Abtahi P, Landry B, Yang J, Pavone M, Follmer S and Landay J A. 2019b. Beyond the force: using quadcopters to appropriate objects and the environment for haptics in virtual reality//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: 1-13 [DOI: 10.1145/3290605.3300589]

Alghofaili R, Sawahata Y, Huang H K, Wang H C, Shiratori T and Yu L F. 2019. Lost in style: gaze-driven adaptive aid for VR navigation//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: 1-12 [DOI: 10.1145/3290605.3300578]

Alper B, Hollerer T, Kuchera-Morin J A and Forbes A. 2011. Stereoscopic highlighting: 2D graph visualization on stereo displays. IEEE Transactions on Visualization and Computer Graphics, 17(12): 2325-2333 [DOI: 10.1109/TVCG.2011.234]

Amoh J and Odame K. 2015. DeepCough: a deep convolutional neural network in a wearable cough detection system//Proceedings of 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS). Atlanta, USA: IEEE: 1-4 [DOI: 10.1109/BioCAS.2015.7348395]

Ando H, Kitahara Y and Hataoka N. 1994. Evaluation of multi-modal interface using spoken language and pointing gesture on interior design system//Proceedings of the 3rd International Conference on Spoken Language Processing. Yokohama, Japan: ISCA: #77

Araujo B, Jota R, Perumal V, Yao J X, Singh K and Wigdor D. 2016. Snake charmer: physically enabling virtual objects//Proceedings of the TEI'16: the 10th International Conference on Tangible, Embedded, and Embodied Interaction. Eindhoven, the Netherlands: ACM: 218-226 [DOI: 10.1145/2839462.2839484]

Arık S Ö, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y G, Li X, Miller J, Ng A, Raiman J, Sengupta S and Shoeybi M. 2017. Deep voice: real-time neural text-to-speech//Proceedings of the 34th International Conference on Machine Learning. Sydney,

Australia: PMLR: 195-204

Baevski A, Schneider S and Auli M. 2020a. Vq-wav2vec: self-supervised learning of discrete speech representations [EB/OL]. [2022-01-20]. https://arxiv.org/pdf/1910.05453v2.pdf

Baevski A, Zhou Y H, Mohamed A and Auli M. 2020b. Wav2vec 2.0: a framework for self-supervised learning of speech representations [EB/OL]. [2022-01-20]. https://arxiv.org/pdf/2006.11477.pdf

Bakhshi A, Wong A S W and Chalup S K. 2020. End-to-end speech emotion recognition based on time and frequency information using deep neural networks//Proceedings of the 24th European Conference on Artificial Intelligence. Santiago de Compostela, Spain: IOS Press: 969-975

Baloup M, Pietrzak T and Casiez G. 2019. RayCursor: a 3D pointing facilitation technique based on raycasting//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #101 [DOI: 10.1145/3290605.3300331]

Barbieri F, Ballesteros M, Ronzano F and Saggion H. 2018. Multi-modal emoji prediction//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA: ACL: 679-686 [DOI: 10.18653/v1/N18-2107]

Bourguet M L. 2003. Designing and prototyping multi-modal commands//Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction. Zurich, Switzerland: IOS Press: 717-720

Büschel W, Chen J, Dachselt R, Drucker S, Dwyer T, Görg C, Isenberg T, Kerren A, North C and Stuerzlinger W. 2018. Interaction for immersive analytics//Marriott K, Schreiber F, Dwyer T, Klein K, Riche N H, Itoh T, Stuerzlinger W and Thomas B H, eds. Immersive Analytics. Cham: Springer: 95-138 [DOI: 10.1007/978-3-030-01388-2_4]

Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket T, Douville B, Prevost S and Stone M. 1994. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents//Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques. New York, USA: ACM: 413-420 [DOI: 10.1145/192161.192272]

Chen C. 2021. Research on Active Fatigue Driving Detection Based on Audio Perception. Tianjin: Tianjin University of Technology (陈超. 2021. 基于音频感知的主动疲劳驾驶检测研究. 天津: 天津理工大学)

Chen C J, Wang Z W, Wu J, Wang X T, Guo L Z, Li Y F and Liu S X. 2021. Interactive graph construction for graph-based semi-supervised learning. IEEE Transactions on Visualization and Computer Graphics, 27(9): 3701-3716 [DOI: 10.1109/TVCG.2021.3084694]

Chen N X, Watanabe S, Villalba J, Żelasko P and Dehak N. 2020a. Non-autoregressive transformer for speech recognition. IEEE Signal Processing Letters, 28: 121-125 [DOI: 10.1109/LSP.2020.3044547]

Chen X Y, Xu J M and Xu B. 2019. A working memory model for task-oriented dialog response generation//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL: 2687-2693 [DOI: 10.18653/v1/P19-1258]

Chen Y Q, Qin X, Wang J D, Yu C H and Gao W. 2020b. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems, 35(4): 83-93 [DOI: 10.1109/MIS.2020.2988604]

Cheng L P, Roumen T, Rantzsch H, Köhler S, Schmidt P, Kovacs R, Jasper J, Kemper J and Baudisch P. 2015. TurkDeck: physical virtual reality based on people//Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology. Charlotte, USA: ACM: 417-426 [DOI: 10.1145/2807442.2807463]

Chi E A, Salazar J and Kirchhoff K. 2021. Align-refine: non-autoregressive speech recognition via iterative realignment//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [s.l.]: Association for Computational Linguistics: 1920-1927 [DOI: 10.18653/v1/2021.naacl-main.154]

Chiu C C and Raffel C. 2018. Monotonic chunkwise attention//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Choi I, Culbertson H, Miller M R, Olwal A and Follmer S. 2017. Grabity: a wearable haptic interface for simulating weight and grasping in virtual reality//Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. Québec City, Canada: ACM: 119-130 [DOI: 10.1145/3126594.3126599]

Choi I, Hawkes E W, Christensen D L, Ploch C J and Follmer S. 2016. Wolverine: a wearable haptic interface for grasping in virtual reality//Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Korea (South): IEEE: 986-993 [DOI: 10.1109/IROS.2016.7759169]

Chu X T, Xie X, Ye S N, Lu H L, Xiao H G, Yuan Z Q, Chen Z T, Zhang H and Wu Y C. 2022. TIVEE: visual exploration and explanation of badminton tactics in immersive visualizations. IEEE Transactions on Visualization and Computer Graphics, 28(1): 118-128 [DOI: 10.1109/TVCG.2021.3114861]

Cordeil M, Bach B, Cunningham A, Montoya B, Smith R T, Thomas B H and Dwyer T. 2020. Embodied axes: tangible, actuated interaction for 3d augmented reality data spaces//Proceedings of 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA: ACM: 1-12 [DOI: 10.1145/3313831.3376613]

Cui C, Wang W J, Song X M, Huang M L, Xu X S and Nie L Q. 2019. User attention-guided multi-modal dialog systems//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris, France: ACM: 445-454 [DOI: 10.1145/3331184.3331226]

Dai Y Y, Jin Y, Ma Y, Yang Z X and Yu J J. 2021. Speech emotion

recognition based on efficient channel attention. Journal of Signal Processing, 37(10): 1835-1842 (戴妍妍, 金赟, 马勇, 杨子秀, 俞佳佳. 2021. 基于高效通道注意力机制的语音情感识别方法. 信号处理, 37(10): 1835-1842) [DOI: 10.16798/j.issn.1003-0530.2021.10.006]

Debie E, Rojas R F, Fidock J, Barlow M, Kasmarik K, Anavatti S, Garratt M and Abbass H A. 2021. Multi-modal fusion for objective assessment of cognitive workload: a review. IEEE Transactions on Cybernetics, 51(3): 1542-1555 [DOI: 10.1109/TCYB.2019.2939399]

Deng Z K, Weng D, Liang Y X, Bao J, Zheng Y, Schreck T, Xu M L and Wu Y C. 2021. Visual cascade analytics of large-scale spatiotemporal data. IEEE Transactions on Visualization and Computer Graphics:#9397369 [DOI: 10.1109/TVCG.2021.3071387]

Deng Z K, Weng D, Xie X, Bao J, Zheng Y, Xu M L, Chen W and Wu Y C. 2022. Compass: towards better causal analysis of urban time series. IEEE Transactions on Visualization and Computer Graphics, 28(1): 1051-1061 [DOI: 10.1109/TVCG.2021.3114875]

Dragicevic P, Jansen Y and Moere A V. 2021. Data physicalization//Vanderdonckt J, Palanque P and Winckler M, eds. Handbook of Human Computer Interaction. Cham: Springer: 1-51 [DOI: 10.1007/978-3-319-27648-9_94-1]

Drogemuller A, Cunningham A, Walsh J, Cordeil M, Ross W and Thomas B. 2018. Evaluating navigation techniques for 3D graph visualizations in virtual reality//Proceedings of 2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA). Konstanz, Germany: IEEE: 1-10 [DOI: 10.1109/BDVA.2018.8533895]

Eric M, Krishnan L, Charette F and Manning C D. 2017. Key-value retrieval networks for task-oriented dialogue//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken, Germany: ACL: 37-49 [DOI: 10.18653/v1/W17-5506]

Eyben F, Wöllmer M and Schuller B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM: 1459-1462 [DOI: 10.1145/1873951.1874246]

Fei Z C, Li Z K, Zhang J C, Feng Y and Zhou J. 2021. Towards expressive communication with internet memes: a new multi-modal conversation dataset and benchmark [EB/OL]. [2022-01-20]. https://arxiv.org/pdf/2109.01839.pdf

Filho J A W, Freitas C M D S and Nedel L. 2019. Comfortable immersive analytics with the VirtualDesk metaphor. IEEE Computer Graphics and Applications, 39(3): 41-53 [DOI: 10.1109/MCG.2019.2898856]

Filho J A W, Stuerzlinger W and Nedel L. 2020. Evaluating an immersive space-time cube geovisualization for intuitive trajectory data exploration. IEEE Transactions on Visualization and Computer Graphics, 26(1): 514-524 [DOI: 10.1109/TVCG.2019.2934415]

Franklin K M and Roberts J C. 2003. Pie chart sonification//Proceedings of the 7th International Conference on Information Visualization, 2003. IV 2003. London, UK: IEEE: 4-9 [DOI: 10.1109/IV.2003.1217949]

Fu L, Li X X, Wang R Y, Fan L, Zhang Z C, Chen M, Wu Y Z and He X D. 2021. SCaLa: supervised contrastive learning for end-to-end speech recognition. [EB/OL]. [2022-01-20]. https://arxiv.org/pdf/2110.04187.pdf

Fujie S, Ejiri Y, Matsusaka Y, Kikuchi H and Kobayashi T. 2003. Recognition of para-linguistic information and its application to spoken dialogue system//Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721). St Thomas, USA: IEEE: 231-236 [DOI: 10.1109/ASRU.2003.1318446]

Fujie S, Yagi D, Matsusaka Y, Kikuchi H and Kobayashi T. 2004. Spoken dialogue system using prosody as para-linguistic information//Proceedings of the Speech Prosody 2004. Nara, Japan: ISCA

Funk M, Müller F, Fendrich M, Shene M, Kolvenbach M, Dobbertin N, Günther S and Mühlhäuser M. 2019. Assessing the accuracy of point and teleport locomotion with orientation indication for virtual reality using curved trajectories//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #147 [DOI: 10.1145/3290605.3300377]

Gannon M, Grossman T and Fitzmaurice G. 2015. Tactum: a skin-centric approach to digital design and fabrication//Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul, Korea(South): ACM: 1779-1788 [DOI: 10.1145/2702123.2702581]

Gannon M, Grossman T and Fitzmaurice G. 2016. ExoSkin: on-body fabrication//Proceedings of 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA: ACM: 5996-6007 [DOI: 10.1145/2858036.2858576]

Gong J, Gupta A and Benko H. 2020. Acustico: surface tap detection and localization using wrist-based acoustic TDOA sensing//Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. New York, USA: ACM: 406-419 [DOI: 10.1145/3379337.3415901]

Gong Y, Chung Y A and Glass J. 2021a. PSLA: improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 3292-3306 [DOI: 10.1109/TASLP.2021.3120633]

Gong Y, Chung Y A and Glass J R. 2021b. AST: audio spectrogram transformer//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 571-575

Goto M, Itou K and Hayamizu S. 2002. Speech completion: on-demand completion assistance using filled pauses for speech input interfaces//Proceedings of the 7th International Conference on Spoken Language Processing. Denver, USA: ISCA: 1489-1492

Goto M, Kitayama K, Itou K and Kobayashi T. 2004. Speech spotter: on-demand speech recognition in human-human conversation on the telephone or in face-to-face situations//Proceedings of the 8th International Conference on Spoken Language Processing. Jeju Island, Korea(South): ISCA: 1533-1536

Goto M, Omoto Y, Itou K and Kobayashi T. 2003. Speech shift: direct speech-input-mode switching through intentional control of voice pitch//Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland: ISCA: 1201-1204

Groeger D and Steimle J. 2017. ObjectSkin: augmenting everyday objects with hydroprinted touch sensors and displays. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(4): #134 [DOI: 10.1145/3161165]

Guo J, Weng D D, Fang H, Zhang Z L, Ping J M, Liu Y and Wang Y T. 2020. Exploring the differences of visual discomfort caused by long-term immersion between virtual environments and physical environments//Proceedings of 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Atlanta, USA: IEEE: 443-452 [DOI: 10.1109/VR46266.2020.00065]

Guo P C, Boyer F, Chang X K, Hayashi T, Higuchi Y, Inaguma H, Kamo N, Li C D, Garcia-Romero D, Shi J T, Shi J, Watanabe S, Wei K, Zhang W Y and Zhang Y K. 2021. Recent developments on espnet toolkit boosted by conformer//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 5874-5878 [DOI: 10.1109/ICASSP39728.2021.9414858]

Gupta A, Irudayaraj A A R and Balakrishnan R. 2017. HapticClench: investigating squeeze sensations using memory alloys//Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. Québec City, Canada: ACM: 109-117 [DOI: 10.1145/3126594.3126598]

Gupta S, Morris D, Patel S and Tan D. 2012. SoundWave: using the Doppler effect to sense gestures//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Austin, USA: ACM: 1911-1914 [DOI: 10.1145/2207676.2208331]

Haber J, Baumgärtner T, Takmaz E, Gelderloos L, Bruni E and Fernández R. 2019. The PhotoBook dataset: building common ground through visually-grounded dialogue//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy: ACL: 1895-1910

Han T, Hasan K, Nakamura K, Gomez R and Irani P. 2017. SoundCraft: enabling spatial interactions on smartwatches using hand generated acoustics//Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. Québec City, Canada: ACM: 579-591 [DOI: 10.1145/3126594.3126612]

Han T, Li J N, Hasan K, Nakamura K, Gomez R, Balakrishnan R and Irani P. 2018. PageFlip: leveraging page-flipping gestures for efficient command and value selection on smartwatches//Proceedings of 2018 CHI Conference on Human Factors in Computing Systems. Montreal, Canada: ACM: #529 [DOI: 10.1145/3173574.3174103]

Han W J, Li H F and Han J Q. 2008. Speech emotion recognition with combined short and long term features. Journal of Tsinghua University (Science and Technology), 48(S1): 708-714 (韩文静, 李海峰, 韩纪庆. 2008. 基于长短时特征融合的语音情感识别方法. 清华大学学报(自然科学版), 48(S1): 708-714) [DOI: 10.16511/j.cnki.qhdxxb.2008.s1.023]

Han W J, Li H F, Ruan H B and Ma L. 2014. Review on speech emotion recognition. Journal of Software, 25(1): 37-50 (韩文静, 李海峰, 阮华斌, 马琳. 2014. 语音情感识别研究进展综述. 软件学报, 25(1): 37-50) [DOI: 10.13328/j.cnki.jos.004497]

Harada S, Landay J A, Malkin J, Li X and Bilmes J A. 2006. The vocal joystick: evaluation of voice-based cursor control techniques//Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility. Portland, USA: ACM: 197-204

Harada S, Wobbrock J O, Malkin J, Bilmes J A and Landay J A. 2009. Longitudinal study of people learning to use continuous voice-based cursor control//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Boston, USA: ACM: 347-356 [DOI: 10.1145/1518701.1518757]

Harrison C, Benko H and Wilson A D. 2011. OmniTouch: wearable multitouch interaction everywhere//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara, USA: ACM: 441-450 [DOI: 10.1145/2047196.2047255]

Harrison C and Faste H. 2014. Implications of location and touch for on-body projected interfaces//Proceedings of 2014 Conference on Designing Interactive Systems. Vancouver, Canada: ACM: 543-552 [DOI: 10.1145/2598510.2598587]

Heo S, Hung C, Lee G and Wigdor D. 2018. Thor's hammer: an ungrounded force feedback device utilizing propeller-induced propulsive force//Proceedings of 2018 CHI Conference on Human Factors in Computing Systems. Montreal, Canada: ACM: #525 [DOI: 10.1145/3173574.3174099]

Hershey S, Chaudhuri S, Ellis D P W, Gemmeke J F, Jansen A, Moore R C, Plakal M, Platt D, Saurous R A, Seybold B, Slaney M, Weiss R J and Wilson K. 2017. CNN architectures for large-scale audio classification//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE: 131-135 [DOI: 10.1109/ICASSP.2017.7952132]

Higuchi Y, Inaguma H, Watanabe S, Ogawa T and Kobayashi T. 2021. Improved mask-CTC for non-autoregressive end-to-end ASR//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 8363-8367 [DOI: 10.1109/ICASSP39728.2021.9414198]

House B, Malkin J and Bilmes J. 2009. The VoiceBot: a voice controlled robot arm//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Boston, USA: ACM: 183-192 [DOI: 10.1145/1518701.1518731]

Hsu W N, Sriram A, Baevski A, Likhomanenko T, Xu Q T, Pratap V, Kahn J, Lee A, Collobert R, Synnaeve G and Auli M. 2021. Robust wav2vec 2.0: analyzing domain shift in self-supervised pre-training//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 721-725

Hu M. 2015. Exploring new paradigms for accessible 3D printed graphs//Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility. Lisbon, Portugal: ACM: 365-366 [DOI: 10.1145/2700648.2811330]

Hu Z M, Bulling A, Li S and Wang G P. 2021. FixationNet: forecasting eye fixations in task-oriented virtual environments. IEEE Transactions on Visualization and Computer Graphics, 27(5): 2681-2690 [DOI: 10.1109/TVCG.2021.3067779]

Huang C C, Gong W, Fu W L and Feng D Y. 2014. Research of speech emotion recognition based on DBNs. Journal of Computer Research and Development, 51(S1): 75-80 (黄晨晨, 巩微, 伏文龙, 冯东煜. 2014. 基于深度信念网络的语音情感识别的研究. 计算机研究与发展, 51(S1): 75-80)

Huang D Y, Chan L W, Yang S, Wang F, Liang R H, Yang S N, Hung Y P and Chen B Y. 2016. DigitSpace: designing thumb-to-fingers touch interfaces for one-handed and eyes-free interactions//Proceedings of 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA: ACM: 1526-1537 [DOI: 10.1145/2858036.2858483]

Huang H Y, Ning C W, Wang P Y, Cheng J H and Cheng L P. 2020a. Haptic-go-round: a surrounding platform for encounter-type haptics in virtual reality experiences//Proceedings of 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA: ACM: 1-10 [DOI: 10.1145/3313831.3376476]

Huang M K, Zhang J, Cai M, Zhang Y, Yao J L, You Y B, He Y and Ma Z J. 2020b. Improving RNN transducer with normalized jointer network [EB/OL]. [2022-01-20]. https://arxiv.org/pdf/2011.01576.pdf

Huang W Y, Hu W C, Yeung Y T and Chen X. 2020c. Conv-transformer transducer: low latency, low frame rate, streamable end-to-end speech recognition//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 5001-5005

Hurter C, Riche N H, Drucker S M, Cordeil M, Alligier R and Vuillemot R. 2019. FiberClay: sculpting three dimensional trajectories to reveal structural insights. IEEE Transactions on Visualization and Computer Graphics, 25(1): 704-714 [DOI: 10.1109/TVCG.2018.2865191]

Igarashi T and Hughes J F. 2001. Voice as sound: using non-verbal voice input for interactive control//Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology. Orlando, USA: ACM: 155-156 [DOI: 10.1145/502348.502372]

Inaguma H, Mimura M and Kawahara T. 2020a. Enhancing monotonic multihead attention for streaming ASR//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 2137-2141

Inaguma H, Mimura M and Kawahara T. 2020b. CTC-synchronous training for monotonic attention model//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 571-575

Ion A, Wang E J and Baudisch P. 2015. Skin drag displays: dragging a physical tactor across the user's skin produces a stronger tactile stimulus than vibrotactile//Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul, Korea (South): ACM: 2501-2504 [DOI: 10.1145/2702123.2702459]

Iravantchi Y, Goel M and Harrison C. 2019. BeamBand: hand gesture sensing with ultrasonic beamforming//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #15 [DOI: 10.1145/3290605.3300245]

Ishii H and Ullmer B. 1997. Tangible bits: towards seamless interfaces between people, bits and atoms//Proceedings of 1997 ACM SIGCHI Conference on Human Factors in Computing Systems. Atlanta, USA: ACM: 234-241 [DOI: 10.1145/258549.258715]

Jansen Y, Isenberg P, Dykes J, Carpendale S, Subramanian S and Keefe D F. 2014. Death of the Desktop Envisioning Visualization without Desktop Computing. Retrieved January, 16: 2017

Je S, Rooney B, Chan L W and Bianchi A. 2017. tactoRing: a skin-drag discrete display//Proceedings of 2017 CHI Conference on Human Factors in Computing Systems. Denver, USA: ACM: 3106-3114 [DOI: 10.1145/3025453.3025703]

Jia Y, Zhang Y, Weiss R J, Wang Q, Shen J, Ren F, Chen Z F, Nguyen P, Pang R M, Moreno I L and Wu Y H. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 4485-4495

Jiang D W, Lei X N, Li W B, Luo N, Hu Y X, Zou W and Li X G. 2019a. Improving transformer-based speech recognition using unsupervised pre-training. [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/1910.09932.pdf

Jiang D W, Li W B, Cao M, Zou W and Li X G. 2021. Speech SimCLR: combining contrastive and reconstruction objective for self-supervised speech representation learning//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 1544-1548

Jiang H Y, Weng D D, Zhang Z L and Chen F. 2019b. HiFinger: one-handed text entry technique for virtual environments based on touches between fingers. Sensors, 19(14): #3063 [DOI: 10.3390/

s19143063]

Jin H J, Holz C and Hornbæk K. 2015. Tracko: ad-hoc mobile 3D tracking using Bluetooth low energy and inaudible signals for cross-device interaction//Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology. Charlotte, USA: ACM: 147-156 [DOI: 10.1145/2807442.2807475]

Jin X C. 2007. A Study on Recognition of Emotions in Speech. Hefei: University of Science and Technology of China (金学成. 2007. 基于语音信号的情感识别研究. 合肥: 中国科学技术大学)

Kitayama K, Goto M, Itou K and Kobayashi T. 2003. Speech starter: noise-robust endpoint detection by using filled pauses//Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland: ISCA: 1237-1240

Kobayashi T and Fujie S. 2013. Conversational robots: an approach to conversation protocol issues that utilizes the paralinguistic information available in a robot-human setting. Acoustical Science and Technology, 34(2): 64-72 [DOI: 10.1250/ast.34.64]

Kong H K, Zhu W J, Liu Z C and Karahalios K. 2019. Understanding visual cues in visualizations accompanied by audio narrations//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #50 [DOI: 10.1145/3290605.3300280]

Kong Q Q, Cao Y, Iqbal T, Wang Y X, Wang W W and Plumbley M D. 2020. PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 2880-2894 [DOI: 10.1109/TASLP.2020.3030497]

Kovacs R, Ofek E, Franco M G, Siu A F, Marwecki S, Holz C and Sinclair M. 2020. Haptic PIVOT: on-demand handhelds in VR//Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. Minneapolis, USA: ACM: 1046-1059 [DOI: 10.1145/3379337.3415854]

Kraus M, Weiler N, Oelke D, Kehrer J, Keim D A and Fuchs J. 2020. The impact of immersion on cluster identification tasks. IEEE Transactions on Visualization and Computer Graphics, 26(1): 525-535 [DOI: 10.1109/TVCG.2019.2934395]

Krekhov A and Krüger J. 2019. Deadeye: a novel preattentive visualization technique based on dichoptic presentation. IEEE Transactions on Visualization and Computer Graphics, 25(1): 936-945 [DOI: 10.1109/TVCG.2018.2864498]

Krekhov A, Cmentowski S, Waschk A and Kruger J. 2020. Deadeye visualization revisited: investigation of preattentiveness and applicability in virtual environments. IEEE Transactions on Visualization and Computer Graphics, 26(1): 547-557 [DOI: 10.1109/TVCG.2019.2934370]

Kwok T C K, Kiefer P, Schinazi V R, Adams B and Raubal M. 2019. Gaze-guided narratives: adapting audio guide content to gaze in virtual and real environments//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #491 [DOI: 10.1145/3290605.3300721]

Kwon O H, Muelder C, Lee K and Ma K L. 2016. A study of layout, rendering, and interaction methods for immersive graph visualization. IEEE Transactions on Visualization and Computer Graphics, 22(7): 1802-1815 [DOI: 10.1109/TVCG.2016.2520921]

Langner R, Satkowski M, Büschel W and Dachselt R. 2021. MARVIS: combining mobile devices and augmented reality for visual data analysis//Proceedings of 2021 CHI Conference on Human Factors in Computing Systems. Yokohama, Japan: ACM: #468 [DOI: 10.1145/3411764.3445593]

Laput G, Xiao R, Chen X, Hudson S E and Harrison C. 2014. Skin buttons: cheap, small, low-powered and clickable fixed-icon laser projectors//Proceedings of the 27th annual ACM symposium on User interface software and technology. Honolulu, USA: ACM: 389-394 [DOI: 10.1145/2642918.2647356]

Lee J and Lee G. 2016. Designing a non-contact wearable tactile display using airflows//Proceedings of the 29th Annual Symposium on User Interface Software and Technology. Tokyo, Japan: ACM: 183-194 [DOI: 10.1145/2984511.2984583]

Lee S P, Cheok A D, James T K S, Debra G P L, Jie C W, Chuang W and Farbiz F. 2006. A mobile pet wearable computer and mixed reality system for human-poultry interaction through the internet. Personal and Ubiquitous Computing, 10(5): 301-317 [DOI: 10.1007/s00779-005-0051-6]

Li F, Wu Y, Xie Y D and Yang S. 2021a. A method for detecting respiratory symptoms based on smartphone audio perception in driving environment. CN, CN112309423A (李凡, 吴玥, 解亚东, 杨松. 2021a. 驾驶环境下基于智能手机音频感知的呼吸道症状检测方法. 中国, CN112309423A)

Li F, Wu Y, Xie Y D and Yang S. 2021b. A method of detecting car driving speed based on smartphone audio perception. CN, CN112230208A (李凡, 吴玥, 解亚东, 杨松. 2021b. 一种基于智能手机音频感知的汽车行驶速度检测方法. 中国, CN112230208A)

Li H Y and Fan L W. 2020. Mapping various large virtual spaces to small real spaces: a novel redirected walking method for immersive VR navigation. IEEE Access, 8: 180210-180221 [DOI: 10.1109/ACCESS.2020.3027985]

Li M, Yang B, Levy J, Stolcke A, Rozgic V, Matsoukas S, Papayiannis C, Bone D and Wang C. 2021. Contrastive unsupervised learning for speech emotion recognition//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6329-6333 [DOI: 10.1109/ICASSP39728.2021.9413910]

Li N L, Kim H J, Shen L Y, Tian F, Han T and Yang X D. 2020. HapLinkage: prototyping haptic proxies for virtual hand tools using linkage mechanism//Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. New York, USA: ACM: 1261-1274 [DOI: 10.1145/3379337.3415812]

Li P C, Song Y, McLoughlin I, Guo W and Dai L R. 2018. An attention pooling based representation learning method for speech emotion recognition//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA: 3087-3091

Lien J, Gillian N, Karagozler M E, Amihood P, Schwesig C, Olson E, Raja H and Poupyrev I. 2016. Soli: ubiquitous gesture sensing with millimeter wave radar. ACM Transactions on Graphics, 35(4): #142 [DOI: 10.1145/2897824.2925953]

Liu A T, Li S W and Lee H Y. 2021. TERA: self-supervised learning of transformer encoder representation for speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 2351-2366 [DOI: 10.1109/TASLP.2021.3095662]

Liu A T, Yang S W, Chi P H, Hsu P C and Lee H Y. 2020. Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 6419-6423 [DOI: 10.1109/ICASSP40776.2020.9054458]

Lu H, Pan W, Lane N D, Choudhury T and Campbell A T. 2009. SoundSense: scalable sound sensing for people-centric applications on mobile phones//Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services. Kraków, Poland: ACM: 165-178 [DOI: 10.1145/1555816.1555834]

Ma J, Wang C L, Shene C K and Jiang J F. 2014. A graph-based interface for visual analytics of 3D streamlines and pathlines. IEEE Transactions on Visualization and Computer Graphics, 20(8): 1127-1140 [DOI: 10.1109/TVCG.2013.236]

Madotto A, Wu C S and Fung P. 2018. Mem2Seq: effectively incorporating knowledge bases into end-to-end task-oriented dialog systems//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL: 1468-1478 [DOI: 10.18653/v1/P18-1136]

Maekawa K. 2004. Production and perception of "paralinguistic" information//Proceedings of the Speech Prosody 2004 Nara. 367-374

Mao W G, He J and Qiu L L. 2016. CAT: high-precision acoustic motion tracking//Proceedings of 22nd Annual International Conference on Mobile Computing and Networking. New York, USA: ACM: 69-81 [DOI: 10.1145/2973750.2973755]

Mao W G, Wang M, Sun W, Qiu L L, Pradhan S and Chen Y C. 2019. RNN-based room scale hand motion tracking//Proceedings of the 25th Annual International Conference on Mobile Computing and Networking. Los Cabos, Mexico: ACM: #38 [DOI: 10.1145/3300061.3345439]

Massie T H and Salisbury J K. 1994. The PHANTOM haptic interface: a device for probing virtual objects//ASME Winter Annual Meeting, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. Chicago, USA: DSC

McNeely W A. 1993. Robotic graphics: a new approach to force feedback for virtual reality//IEEE Virtual Reality Annual International Symposium. Seattle, USA: IEEE: 336-341 [DOI: 10.1109/VRAIS.1993.380761]

Munzner T. 2014. Visualization Analysis and Design. CRC Press

Nandakumar R, Iyer V, Tan D and Gollakota S. 2016. FingerIO: using active sonar for fine-grained finger tracking//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA: ACM: 1515-1525 [DOI: 10.1145/2858036.2858580]

Olberding S, Wessely M and Steimle J. 2014. PrintScreen: fabricating highly customizable thin-film touch-displays//Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology. Honolulu, USA: ACM: 281-290 [DOI: 10.1145/2642918.2647413]

Ono M, Shizuki B and Tanaka J. 2013. Touch and activate: adding interactivity to existing objects using active acoustic sensing//The 26th Annual ACM Symposium on User Interface Software and Technology. St. Andrews, Scotland: ACM: 31-40 [DOI: 10.1145/2501988.2501989]

Pan Z G, Gao J L, Wang R N, Yuan Q S, Fan R and She L. 2021. Digital twin registration technique of spatial augmented reality for tangible interaction. Journal of Computer-Aided Design and Computer Graphics, 33(5): 655-661 (潘志庚, 高嘉利, 王若楠, 袁庆曙, 范然, 佘莉. 2021. 面向实物交互的空间增强现实数字孪生法配准技术. 计算机辅助设计与图形学学报, 33(5): 655-661) [DOI: 10.3724/SP.J.1089.2021.18556]

Park J H, Nadeem S, Boorboor S, Marino J and Kaufman A. 2021. CMed: crowd analytics for medical imaging data. IEEE Transactions on Visualization and Computer Graphics, 27(6): 2869-2880 [DOI: 10.1109/TVCG.2019.2953026]

Patnaik B, Batch A and Elmqvist N. 2019. Information olfactation: harnessing scent to convey data. IEEE Transactions on Visualization and Computer Graphics, 25(1): 726-736 [DOI: 10.1109/TVCG.2018.2865237]

Pavlovic V I, Berry G A and Huang T S. 1997. Integration of audio/visual information for use in human-computer intelligent interaction//Proceedings of International Conference on Image Processing. Santa Barbara, USA: IEEE: 121-124 [DOI: 10.1109/ICIP.1997.647399]

Peng C Y, Shen G B, Zhang Y G, Li Y L and Tan K. 2007. BeepBeep: a high accuracy acoustic ranging system using COTS mobile devices//Proceedings of the 5th International Conference on Embedded Networked Sensor Systems. Sydney, Australia: ACM: 1-14 [DOI: 10.1145/1322263.1322265]

Pepino L, Riera P and Ferrer L. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 3400-3404

Prouzeau A, Cordeil M, Robin C, Ens B, Thomas B H and Dwyer T.

2019. Scaptics and highlight-planes: immersive interaction techniques for finding occluded features in 3D scatterplots//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #325 [DOI: 10.1145/3290605.3300555]

Qin L B, Liu Y J, Che W X, Wen H Y, Li Y M and Liu T. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL: 133-142 [DOI: 10.18653/v1/D19-1013]

Qin L B, Xu X, Che W X, Zhang Y and Liu T. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [s. l.]: ACL: 6344-6354

Qin Y, Yu C, Li Z H, Zhong M Y, Yan Y K and Shi Y C. 2021. ProxiMic: convenient voice activation via close-to-mic speech detected by a single microphone//Proceedings of 2021 CHI Conference on Human Factors in Computing Systems. Yokohama, Japan: ACM: #8 [DOI: 10.1145/3411764.3445687]

Renner R S, Velichkovsky B M and Helmert J R. 2013. The perception of egocentric distances in virtual environments——A review. ACM Computing Surveys, 46(2): #23 [DOI: 10.1145/2543581.2543590]

Röddiger T, Clarke C, Wolffram D, Budde M and Beigl M. 2021. EarRumble: discreet hands-and eyes-free input by voluntary tensor tympani muscle contraction//Proceedings of 2021 CHI Conference on Human Factors in Computing Systems. Yokohama, Japan: ACM: #743 [DOI: 10.1145/3411764.3445205]

Rossi M, Feese S, Amft O, Braune N, Martis S and Tröster G. 2013. AmbientSense: a real-time ambient sound recognition system for smartphones//Proceedings of 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). San Diego, USA: IEEE: 230-235 [DOI: 10.1109/PerComW.2013.6529487]

Ruan W J, Sheng Q Z, Yang L, Gu T, Xu P P and Shangguan L F. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal//Proceedings of 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg, Germany: ACM: 474-485 [DOI: 10.1145/2971648.2971736]

Saakes D, Yeo H S, Noh S T, Han G and Woo W. 2016. Mirror mirror: an on-body t-shirt design system//Proceedings of 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA: ACM: 6058-6063 [DOI: 10.1145/2858036.2858282]

Sadhu S, He D, Huang C W, Mallidi S H, Wu M H, Rastrow A, Stolcke A, Droppo J and Maas R. 2021. Wav2vec-C: a self-supervised model for speech representation learning//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 711-715

Sainath T N, Pang R M, Rybach D, He Y Z, Prabhavalkar R, Li W, Visontai M, Liang Q, Strohman T, Wu Y H, McGraw I and Chiu C C. 2019. Two-pass end-to-end speech recognition [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/1908.10992.pdf

Saponas T S, Tan D S, Morris D, Balakrishnan R, Turner J and Landay J A. 2009. Enabling always-available input with muscle-computer interfaces//Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology. Victoria, Canada: ACM: 167-176 [DOI: 10.1145/1622176.1622208]

Satt A, Rozenberg S and Hoory R. 2017. Efficient emotion recognition from speech using deep learning on spectrograms//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA: 1089-1093

Schneider S, Baevski A, Collobert R and Auli M. 2019. Wav2vec: unsupervised pre-training for speech recognition//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 3465-3469

Schrapel M, Stadler M L and Rohs M. 2018. Pentelligence: combining pen tip motion and writing sounds for handwritten digit recognition//Proceedings of 2018 CHI Conference on Human Factors in Computing Systems. Montreal, Canada: ACM: #131 [DOI: 10.1145/3173574.3173705]

Shen J, Pang R M, Weiss R J, Schuster M, Jaitly N, Yang Z H, Chen Z F, Zhang Y, Wang Y X, Skerrv-Ryan R, Saurous R A, Agiomvrgiannakis Y and Wu Y H. 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 373-376 [DOI: 10.1109/ICASSP.2018.8461368]

Shigeyama J, Hashimoto T, Yoshida S, Narumi T, Tanikawa T and Hirose M. 2019. Transcalibur: a weight shifting virtual reality controller for 2D shape rendering based on computational perception model//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: #11 [DOI: 10.1145/3290605.3300241]

Sidenmark L, Clarke C, Zhang X S, Phu J and Gellersen H. 2020. Outline pursuits: gaze-assisted selection of occluded objects in virtual reality//Proceedings of 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA: ACM: 1-13 [DOI: 10.1145/3313831.3376438]

Siu A F, Sinclair M, Kovacs R, Ofek E, Holz C and Cutrell E. 2020. Virtual reality without vision: a haptic and auditory white cane to navigate complex virtual worlds//Proceedings of 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA: ACM Press: 1-13 [DOI: 10.1145/3313831.3376353]

Sotelo J, Mehri S, Kumar K, Santos J F, Kastner K, Courville A C and Bengio Y. 2017. Char2 Wav: end-to-end speech synthesis//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Ssin S Y, Walsh J A, Smith R T, Cunningham A and Thomas B H.

2019. GeoGate: correlating geo-temporal datasets using an augmented reality space-time cube and tangible interactions//Proceedings of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Osaka, Japan: IEEE: 210-219 [DOI: 10.1109/VR.2019.8797812]

Sun L C, Liu B, Tao J H and Lian Z. 2021. Multi-modal cross-and self-attention network for speech emotion recognition//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 4275-4279 [DOI: 10.1109/ICASSP39728.2021.9414654]

Suzuki R, Hedayati H, Zheng C, Bohn J L, Szafir D, Do E Y L, Gross M D and Leithinger D. 2020. RoomShift: room-scale dynamic haptics for VR with furniture-moving swarm robots//Proceedings of 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA: ACM: 1-11 [DOI: 10.1145/3313831.3376523]

Thomaz E, Zhang C, Essa I and Abowd G D. 2015. Inferring meal eating activities in real world settings from ambient sounds: a feasibility study. IUI, 2015: 427-431 [DOI: 10.1145/2678025.2701405]

Tian Y, Yao H T, Cai M, Liu Y M and Ma Z J. 2021a. Improving RNN transducer modeling for small-footprint keyword spotting//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 5624-5628 [DOI: 10.1109/ICASSP39728.2021.9414339]

Tian Z K, Yi J Y, Bai Y, Tao J H, Zhang S and Wen Z Q. 2020. One in a hundred: Select the best predicted sequence from numerous candidates for streaming speech recognition [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/2010.14791.pdf

Tian Z K, Yi J Y, Bai Y, Tao J H, Zhang S and Wen Z Q. 2021b. FSR: accelerating the inference process of transducer-based models by applying fast-skip regularization//Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 4034-4038

Tian Z K, Yi J Y, Tao J H, Bai Y and Wen Z Q. 2019. Self-attention transducers for end-to-end speech recognition//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 4395-4399

Tzirakis P, Zhang J H and Schuller B W. 2018. End-to-end speech emotion recognition using deep neural networks//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 5089-5093 [DOI: 10.1109/ICASSP.2018.8462677]

Usher W, Klacansky P, Federer F, Bremer P T, Knoll A, Yarch J, Angelucci A and Pascucci V. 2018. A virtual reality visualization tool for neuron tracing. IEEE Transactions on Visualization and Computer Graphics, 24(1): 994-1003 [DOI: 10.1109/TVCG.2017.2744079]

Valin J M and Skoglund J. 2019. LPCNET: improving neural speech synthesis through linear prediction//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 4384-4289 [DOI: 10.1109/ICASSP.2019.8682804]

van den Oord A, Dieleman S, Zen H G, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A W and Kavukcuoglu K. 2016. WaveNet: a generative model for raw audio//Proceedings of the 9th ISCA Speech Synthesis Workshop. Sunnyvale, USA: ISCA: #125

Wagner J, Stuerzlinger W and Nedel L. 2021. Comparing and combining virtual hand and virtual ray pointer interactions for data manipulation in immersive analytics. IEEE Transactions on Visualization and Computer Graphics, 27(5): 2513-2523 [DOI: 10.1109/TVCG.2021.3067759]

Wang A R and Gollakota S. 2019. MilliSonic: pushing the limits of acoustic motion tracking//Proceedings of 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK: ACM: 1-11 [DOI: 10.1145/3290605.3300248]

Wang C Y, Hsiu M C, Chiu P T, Chang C H, Chan L W, Chen B Y and Chen M Y. 2015. PalmGesture: using palms as gesture interfaces for eyes-free input//Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services. Copenhagen, Denmark: ACM: 217-226 [DOI: 10.1145/2785830.2785885]

Wang C Y, Liu J, Chen Y Y, Liu H B, Xie L, Wang W, He B B and Lu S L. 2018. Multi-touch in the air: device-free finger tracking and gesture recognition via COTS RFID//Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications. Honolulu, USA: IEEE: 1691-1699 [DOI: 10.1109/INFOCOM.2018.8486346]

Wang H, Zhang D Q, Wang Y S, Ma J Y, Wang Y X and Li S J. 2017a. RT-Fall: a real-time and contactless fall detection system with commodity WiFi devices. IEEE Transactions on Mobile Computing, 16(2): 511-526 [DOI: 10.1109/TMC.2016.2557795]

Wang J D, Chen Y Q, Hao S J, Peng X H and Hu L S. 2019. Deep learning for sensor-based activity recognition: a survey [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/1707.03502.pdf

Wang S, Zhu D, Yu H, and Wu Y D. 2020a. Immersive WYSIWYG (what you see is what you get) volume visualization//2020 IEEE Pacific Visualization Symposium (PacificVis). Tianjin, China: IEEE: 166-170 [DOI: 10.1109/PacificVis48177.2020.1001]

Wang T, Tao J H, Fu R B, Yi J Y, Wen Z Q and Zhong R X. 2020b. Spoken content and voice factorization for few-shot speaker adaptation//Proceedings of the 21st Annual Conference of the International Speech Communication Association, Virtual Event. Shanghai, China: ISCA: 796-800

Wang W, Liu A X and Sun K. 2016. Device-free gesture tracking using acoustic signals: demo//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. New York, USA: ACM: 497-498 [DOI: 10.1145/2973750.2987385]

Wang Y F, Peng T Q, Lu H H, Wang H R, Xie X, Qu H M and Wu Y

C. 2022a. Seek for success: a visualization approach for understanding the dynamics of academic careers. IEEE Transactions on Visualization and Computer Graphics, 28(1): 475-485 [DOI: 10.1109/TVCG.2021.3114790]

Wang Y T, Ding J X, Chatterjee I, Parizi F S, Zhuang Y Z, Yan Y K, Patel S and Shi Y C. 2022b. FaceOri: tracking head position and orientation using ultrasonic ranging on earphones//Proceedings of 2022 CHI Conference on Human Factors in Computing Systems (CHI'22). New York, USA: ACM: 1-12

Wang Y W, Lin Y H, Ku P S, Miyatake Y, Mao Y H, Chen P Y, Tseng C M and Chen M Y. 2021. JetController: high-speed ungrounded 3-DoF force feedback controllers using air propulsion jets//Proceedings of 2021 CHI Conference on Human Factors in Computing Systems. Yokohama, Japan: ACM: #124 [DOI: 10.1145/3411764.3445549]

Wang Y X, Skerry-Ryan R J, Stanton D, Wu Y H, Weiss R J, Jaitly N, Yang Z H, Xiao Y, Chen Z F, Bengio S, Le Q V, Agiomyrgiannakis Y, Clark R and Saurous R A. 2017b. Tacotron: towards end-to-end speech synthesis//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA: 4006-4010

Ward J A, Lukowicz P and Tröster G. 2005. Gesture spotting using wrist worn microphone and 3-axis accelerometer//Proceedings of 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies. Grenoble, France: ACM: 99-104 [DOI: 10.1145/1107548.1107578]

Wei W Z and He Q B. 2018. Research on ultrasound-based gesture recognition device. Machinery and Electronics, 36(5): 54-57, 61 (魏文钊, 何清波. 2018. 基于超声波的手势识别设备的研究. 机械与电子, 36(5): 54-57, 61) [DOI: 10.3969/j.issn.1001-2257.2018.05.012]

Weigel M, Mehta V and Steimle J. 2014. More than touch: understanding how people use skin as an input surface for mobile computing//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Toronto, Canada: ACM: 179-188 [DOI: 10.1145/2556288.2557239]

Weigel M and Steimle J. 2017. DeformWear: deformation input on tiny wearable devices. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(2): #28 [DOI: 10.1145/3090093]

Withana A, Groeger D and Steimle J. 2018. Tacttoo: a thin and feel-through tattoo for on-skin tactile output//Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. Berlin, Germany: ACM: 365-378 [DOI: 10.1145/3242587.3242645]

Wu C S, Socher R and Xiong C M. 2019. Global-to-local memory pointer networks for task-oriented dialogue//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Xi H W and Kelley A. 2015. Sonification of time-series data sets. Bulletin of the American Physical Society, 60(3)

Xiao R, Cao T, Guo N, Zhuo J, Zhang Y and Harrison C. 2018. LumiWatch: on-arm projected graphics and touch input//Proceedings of 2018 CHI Conference on Human Factors in Computing Systems. Montreal: ACM: #95 [DOI: 10.1145/3173574.3173669]

Xie L, Sheng B, Tan C C, Han H, Li Q and Chen D X. 2010. Efficient tag identification in mobile RFID systems//Proceedings of 2010 Proceedings IEEE INFOCOM. San Diego, USA: IEEE: 1-9 [DOI: 10.1109/INFCOM.2010.5461949]

Xue Y Q, Weng D D, Jiang H Y and Gao Q. 2019. MMRPet: modular mixed reality pet system based on passive props//Proceedings of the 14th Chinese Conference on Image and Graphics Technologies. Beijing, China: Springer: 645-658 [DOI: 10.1007/978-981-13-9917-6_61]

Yamamoto R, Song E and Kim J M. 2020. Parallel wavegan: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 6199-6203 [DOI: 10.1109/ICASSP40776.2020.9053795]

Yan Y K, Yu C, Shi Y T and Xie M X. 2019. PrivateTalk: activating voice input with hand-on-mouth gesture detected by Bluetooth earphones//Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. New Orleans, USA: ACM: 1013-1020 [DOI: 10.1145/3332165.3347950]

Yang D Q, Zhang D Q, Zheng V W and Yu Z Y. 2015. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(1): 129-142 [DOI: 10.1109/TSMC.2014.2327053]

Yang Y L, Cordeil M, Beyer J, Dwyer T, Marriott K and Pfister H. 2021a. Embodied navigation in immersive abstract data visualization: is overview + detail or zooming better for 3D scatterplots? IEEE Transactions on Visualization and Computer Graphics, 27(2): 1214-1224 [DOI: 10.1109/TVCG.2020.3030427]

Yang Y L, Dwyer T, Jenny B, Marriott K, Cordeil M and Chen H H. 2019. Origin-destination flow maps in immersive environments. IEEE Transactions on Visualization and Computer Graphics, 25(1): 693-703 [DOI: 10.1109/TVCG.2018.2865192]

Yang Y L, Dwyer T, Marriott K, Jenny B and Goodwin S. 2021b. Tilt map: interactive transitions between choropleth map, prism map and bar chart in immersive environments. IEEE Transactions on Visualization and Computer Graphics, 27(12): 4507-4519 [DOI: 10.1109/TVCG.2020.3004137]

Yao L N, Ou J F, Cheng C Y, Steiner H, Wang W, Wang G Y and Ishii H. 2015. bioLogic: natto cells as nanoactuators for shape changing interfaces//Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul, Korea (South): ACM: 1-10 [DOI: 10.1145/2702123.2702611]

Yao Z Y, Wu D, Wang X, Zhang B B, Yu F, Yang C, Peng Z D, Chen C Y, Xie L and Lei X. 2021. WeNet: production oriented streaming and non-streaming end-to-end speech recognition toolkit [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/2102.01547.pdf

Ye S N, Chen Z T, Chu X T, Wang Y F, Fu S W, Shen L J, Zhou K and Wu Y C. 2021. ShuttleSpace: exploring and analyzing movement trajectory in immersive visualization. IEEE Transactions on Visualization and Computer Graphics, 27(2): 860-869 [DOI: 10.1109/TVCG.2020.3030392]

Ye S N, Chu X T and Wu Y C. 2021. A survey on immersive visualization. Journal of Computer-Aided Design and Computer Graphics, 33(4): 497-507 (叶帅男, 储向童, 巫英才. 2021. 沉浸式可视化综述. 计算机辅助设计与图形学学报, 33(4): 497-507) [DOI: 10.3724/SP.J.1089.2021.18809]

Yeh C F, Mahadeokar J, Kalgaonkar K, Wang Y Q, Le D, Jain M, Schubert K, Fuegen C and Seltzer M L. 2019. Transformer-transducer: end-to-end speech recognition with self-attention [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/1910.12977.pdf

Yi X, Yu C, Xu W J, Bi X J and Shi Y C. 2017. COMPASS: rotational keyboard on non-touch smartwatches//Proceedings of 2017 CHI Conference on Human Factors in Computing Systems. Denver, USA: ACM: 705-715 [DOI: 10.1145/3025453.3025454]

Yoon S, Byun S, Dey S and Jung K. 2019. Speech emotion recognition using multi-hop attention mechanism//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 2822-2826 [DOI: 10.1109/ICASSP.2019.8683483]

Yoon S, Dey S, Lee H and Jung K. 2020. Attentive modality hopping mechanism for speech emotion recognition//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 3362-3366 [DOI: 10.1109/ICASSP40776.2020.9054229]

Yu J J, Jin Y, Ma Y, Jiang F Z and Dai Y Y. 2021. Emotion recognition from raw speech based on Sinc-Transformer model. Journal of Signal Processing, 37(10): 1880-1888 (俞佳佳, 金赟, 马勇, 姜芳艽, 戴妍妍. 2021. 基于 Sinc-Transformer 模型的原始语音情感识别. 信号处理, 37(10): 1880-1888) [DOI: 10.16798/j.issn.1003-0530.2021.10.011]

Yu T, Jin H M and Nahrstedt K. 2016. WritingHacker: audio based eavesdropping of handwriting via mobile devices//Proceedings of 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg, Germany: ACM: 463-473 [DOI: 10.1145/2971648.2971681]

Yun S, Chen Y C, Mao W G and Qiu L L. 2015. Demo: turning a mobile device into a mouse in the air//Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services. Florence, Italy: ACM: #469

Zen H G, Tokuda K and Black A W. 2009. Statistical parametric speech synthesis. Speech Communication, 51(11): 1039-1064 [DOI: 10.1016/j.specom.2009.04.004]

Zhang B B, Wu D, Yao Z Y, Wang X, Yu F, Yang C, Guo L Y, Hu Y G, Xie L and Lei X. 2020b. Unified streaming and non-streaming two-pass end-to-end model for speech recognition [EB/OL]. [2022-01-26]. https://arxiv.org/pdf/2012.05481.pdf

Zhang C, Bedri A K, Reyes G, Bercik B, Inan O T, Starner T E and Abowd G D. 2016. TapSkin: recognizing on-skin input for smartwatches//Proceedings of 2016 ACM International Conference on Interactive Surfaces and Spaces. Niagara Falls, Canada: ACM: 13-22 [DOI: 10.1145/2992154.2992187]

Zhang C, Waghmare A, Kundra P, Pu Y M, Gilliland S, Ploetz T, Starner T E, Inan O T and Abowd G D. 2017a. FingerSound: recognizing unistroke thumb gestures using a ring. Proceedings of 2017 ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(3): #120 [DOI: 10.1145/3130985]

Zhang C, Xue Q Y, Waghmare A, Jain S, Pu Y M, Hersek S, Lyons K, Cunefare K A, Inan O T and Abowd G D. 2017b. SoundTrak: continuous 3D tracking of a finger using active acoustics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(2): #30 [DOI: 10.1145/3090095]

Zhang C, Xue Q Y, Waghmare A, Meng R C, Jain S, Han Y Z, Li X Y, Cunefare K, Ploetz T, Starner T and Inan O. 2018. FingerPing: recognizing fine-grained hand poses using active acoustic on-body sensing//Proceedings of 2018 CHI Conference on Human Factors in Computing Systems. Montreal, Canada: ACM: #437 [DOI: 10.1145/3173574.3174011]

Zhang Q, Lu H, Sak H, Tripathi A, McDermott E, Koo S and Kumar S. 2020a. Transformer transducer: a streamable speech recognition model with transformer encoders and RNN-T loss//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 7829-7833 [DOI: 10.1109/ICASSP40776.2020.9053896]

Zhang R X, Wu H W, Li W B, Jiang D W, Zou W and Li X G. 2021. Transformer based unsupervised pre-training for acoustic representation learning//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6933-6937 [DOI: 10.1109/ICASSP39728.2021.9414996]

Zhang X T, Fang G X, Dai C K, Verlinden J, Wu J, Whiting E and Wang C C L. 2017c. Thermal-comfort design of personalized casts//The 30th Annual ACM Symposium on User Interface Software and Technology. Québec City, Cananda: ACM: 243-254 [DOI: 10.1145/3126594.3126600]

Zhang Y, Wang D X, Wang Z Q, Zhang Y R and Xiao J. 2019. Passive force-feedback gloves with joint-based variable impedance using layer jamming. IEEE Transactions on Haptics, 12(3): 269-280 [DOI: 10.1109/TOH.2019.2908636]

Zhang Y, Zhou J H, Laput G and Harrison C. 2016. SkinTrack: using the body as an electrical waveguide for continuous finger tracking on

the skin//Proceedings of 2016 CHI Conference on Human Factors in Computing Systems. San Jose, USA: ACM: 1491-1503 [DOI: 10.1145/2858036.2858082]

Zhang Y K, Sun S N and Ma L. 2021. Tiny transducer: a highly-efficient speech recognition model on edge devices//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6024-6028 [DOI: 10.1109/ICASSP39728.2021.9413854]

Zhao J H, Lin Y X and Yuan Z Y. 2021. Designing and simulation of electromagnetic force feedback model focusing on virtual interventional surgery. Journal of Computer-Aided Design and Computer Graphics, 33(8): 1254-1263 (赵俭辉, 林远轩, 袁志勇. 2021. 面向虚拟介入手术的电磁力反馈模型的设计与仿真. 计算机辅助设计与图形学学报, 33(8): 1254-1263) [DOI: 10.3724/SP.J.1089.2021.18703]

Zhao J M, Li R C, Chen S Z and Jin Q. 2018. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions//Proceedings of 2018 on Audio/Visual Emotion Challenge and Workshop. Seoul, Korea (South): ACM: 65-72 [DOI: 10.1145/3266302.3266313]

Zhao L, Jiang C H, Zou C R and Wu Z Y. 2004. A study on emotional feature analysis and recognition in speech. Acta Electronica Sinica, 32(4): 606-609 (赵力, 将春辉, 邹采荣, 吴镇扬. 2004. 语音信号中的情感特征分析和识别的研究. 电子学报, 32(4): 606-609) [DOI: 10.3321/j.issn:0372-2112.2004.04.018]

Zhao L, Liu Y and Song W T. 2021. Tactile perceptual thresholds of electrovibration in VR. IEEE Transactions on Visualization and Computer Graphics, 27(5): 2618-2626 [DOI: 10.1109/TVCG.2021.3067778]

Zhao Y W, Kim L H, Wang Y, Le Goc M and Follmer S. 2017. Robotic assembly of haptic proxy objects for tangible interaction and virtual reality//Proceedings of 2017 ACM International Conference on Interactive Surfaces and Spaces. Brighton, UK: ACM: 82-91 [DOI: 10.1145/3132272.3134143]

Zhou F, Duh H B L and Billinghurst M. 2008. Trends in augmented reality tracking, interaction and display: a review of ten years of ISMAR//The 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. Cambridge, UK: IEEE: 193-202 [DOI: 10.1109/ISMAR.2008.4637362]

Zhou J H, Zhang Y, Laput G and Harrison C. 2016. AuraSense: enabling expressive around-smartwatch interactions with electric field sensing//The 29th Annual Symposium on User Interface Software and Technology. Tokyo, Japan: ACM: 81-86 [DOI: 10.1145/2984511.2984568]

Zhuang Y Z, Wang Y T, Yan Y K, Xu X H and Shi Y C. 2021. ReflecTrack: enabling 3D acoustic position tracking using commodity dual-microphone smartphones//The 34th Annual ACM Symposium on User Interface Software and Technology. New York, USA: ACM: 1050-1062 [DOI: 10.1145/3472749.3474805]

Zhuge M C, Gao D H, Fan D P, Jin L B, Chen B, Zhou H M, Qiu M H and Shao L. 2021. Kaleido-BERT: vision-language pre-training on fashion domain//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12647-12657 [DOI: DOI: 10.1109/CVPR46437.2021.01246]

作者简介

陶建华，1972 年生，男，研究员，主要研究方向为自然口语语音交互、情感计算。
E-mail：jhtao@nlpr.ia.ac.cn

巫英才，男，教授，主要研究方向为可视分析。
E-mail：ycwu@zju.edu.cn

喻纯，男，副教授，主要研究方向为人机交互。
E-mail：chunyu@tsinghua.edu.cn

翁冬冬，男，研究员，主要研究方向为虚拟现实、增强现实、人机交互与数字人。E-mail：crgj@bit.edu.cn

李冠君，男，助理研究员，主要研究方向为人机对话交互。
E-mail：guanjun.li@nlpr.ia.ac.cn

韩腾，男，副研究员，主要研究方向为人机交互、智能感知、触觉反馈。E-mail：hanteng@iscas.ac.cn

王运涛，男，助理研究员，主要研究方向为人机交互、普适计算、生理计算。E-mail：yuntaowang@tsinghua.edu.cn

刘斌，男，副研究员，主要研究方向为情感计算、多模态交互。
E-mail：liubin@nlpr.ia.ac.cn