元宇宙终端:虚拟(增强)现实关键硬科技发展趋势

王立军,李争平,李颖,侯耀辉,王京亮,汪闯,徐志平,贾可豪,刘宇宁,马唯植

北方工业大学信息学院,北京 100144

摘要 元宇宙是第三次互联网革命(Web3.0)的一个战略制高点,头戴式设备(XR)是其主要的接入和交互终端。从微显示、光学系统和感知交互3个核心技术领域,论述了元宇宙终端硬件的关键技术及其发展趋势。在微显示技术方面,重点涉及硅基OLED 和 MicroLED 2种新型显示技术及其产业化;在光学系统方面,主要论述了表面浮雕光栅光波导、体全息光波导和超表面光波导技术;在感知交互方面,着重讨论了手势动作识别和脑电信号识别交互技术。

关键词 元宇宙;虚拟现实(VR);增强现实(AR);微显示;感知交互;人工智能

随着科技的不断发展,虚拟现实与增强现实技术成为新的热点领域。这些技术已改变人们的生活方式,为各个行业带来了深刻影响。虚拟现实技术是新一代信息技术的集大成者,是高度跨学科融合形成的重大前沿技术领域,是新一轮技术和产业变革的战略重点。随着虚拟现实技术的进步、5G网络的普及和元宇宙概念的兴起,全球虚拟现实产业正进入新一轮爆发期。

虚拟现实(virtual reality,VR)是以计算机科学和人机交互技术为核心,生成在视觉、听觉和触觉等方面高度逼真的数字化环境,用户借助必要的穿戴式设备(或者裸眼),与数字化环境中的虚拟对象和空间进行自然交互和相互协作,从而产生身临其境的认知体验和虚实交互作用。从广义上讲,虚拟现实技术包括近年来发展迅速的增强现实(augmented reality,AR)技术和混合现实(mixed reality,MR)技术,这两者有时也统称为扩展现实(extend reality,XR)[1]。

VR一般是创造和用户所在物理世界视觉上完全隔离的虚拟世界,主要应用于消费者领域,如游戏、影视、社交、娱乐和教育培训等,可以为用户提供沉浸式体验。AR和MR则强调虚拟世界和物理世界的融合与交互,即将数字信息与现实世界相融合,创造出更为丰富的交互体验。AR和MR类似,只是在实现技术细节和应用对象上略有不同,一般统称为AR,主要应用于工业领域。

收稿日期:2023-02-22;修回日期:2023-04-26

作者简介:王立军,教授,研究方向为虚拟现实技术,电子信箱:1094029730@qq.com

引用格式:王立军, 李争平, 李颖, 等. 元宇宙终端:虚拟(增强)现实关键硬科技发展趋势[J]. 科技导报, 2023, 41(15): 46-60; doi: 10.3981/j.issn.1000-7857.2023.15.005

1　研究现状

随着计算机科学的发展,特别是计算机图形学、高性能计算和人机自然交互技术的快速发展,虚拟现实是人类模拟现实世界和创造未来虚拟世界的最高水平[2],是高度集成的多学科创新技术。虚拟和增强现实(VR/AR)硬件、软件、内容和应用系统几乎涵盖了下一代信息技术的所有领域。

1.1　虚拟(增强)现实微显示关键技术

微型显示屏幕是VR/AR等硬件实现交互的基础,也是进入"元宇宙"的核心技术之一。相比于手机、平板等电子设备,用于VR/AR的微显示屏有更高的性能需求,包括尺寸、显示亮度、画面密度及产品能耗等指标要求明显提高。目前VR领域最主流的显示技术为硅基有机发光半导体(OLED),具有超薄、高亮度和对比度的硅基OLED微显示器件采用单晶硅晶圆(wafer)为背板,具有自发光、宽视角、高分辨率、低响应时间和低功耗等优点,特别适合应用于近眼显示设备。微发光二极管(MicroLED)是一种全新的显示技术,近年来吸引了越来越多的关注。将微小尺寸和高密度的发光二极管(LED)矩阵集成在芯片上,使得每个像素都可以独立驱动。它具有更高的对比度、更高亮度、更迅速的反应时、使用寿命长、高动态范围等优点,被认为是VR/AR微型显示器的最佳解决方案[3]。

1.2　虚拟(增强)现实光学系统关键技术

头戴显示是VR/AR技术的首要信息呈现设备,其中VR头戴显示屏蔽了真实环境,将计算机生成的虚拟场景以大视场沉浸式的效果呈现给用户;而AR头戴显示则采用透视式光学系统,将计算机生成的增强信息和真实世界融为一体展现给用户。VR和AR头戴显示光学系统通常要求大视场角、高分辨率、真彩色显示;为了使用方便,要求具有长出瞳距离(方便戴眼镜的用户)和大出瞳直径(适应不同瞳距的用户);而"头戴"的使用属性又对系统的体积、重量和显示舒适度要求严苛。上述各种要求互相矛盾、互相制约,使得VR和AR头戴显示系统的设计和制造成为一个极为棘手的问题。

波导是目前最佳的AR眼镜方案,大致分为几何光波导(geometric waveguide)和衍射光波导(diffractive waveguide)2种类型。衍射光波导根据光耦合器的不同可分为采用纳米压印加工技术制备的表面浮雕光栅波导(surface relief grating,SRG)和基于全息干涉光刻技术制备的全息体光栅波导(volumetric holographic grating,VHG)[4]。其中,微软、Magic Leap等多家AR生产企业的规模量产证明了SRG这一技术路线在经济成本上的可行性。全息体光栅由于受到可利用材料的限制,致使其在视场角、光效率、清晰度等方面尚未达到表面浮雕光栅的水平,但因其在量产经济性等方面的发展潜力,业界对此方向的探索未曾停歇。

1.3　虚拟(增强)现实感知交互关键技术

随着VR/AR大范围推广应用,随时随地的自然人机交互成为产业界所关注的焦点,视觉、听觉等多模态智能化交互是技术发展的必然趋势。但是面对智能物联、智能制造、智能感知等领域需求,现有的信息呈现与交互的理论和技术框架无法满足高逼真、高精度、强舒适等需求,主要体现在:

1)目前的VR/AR呈现依然以真实或者虚拟场景为主,融合另一类场景信息,尚未实现"虚""实"场景中人机物信息有机融合,无法满足虚实空间信息的同步变化与实时更新需求。

2)VR/AR中多元要素相互联结、和谐共生,交互内容和交互行为呈现多维且复杂特征,传统呈现与交互环境下的交互方法不适用,特征与规律性的基础理论匮乏。

3)VR/AR交互涉及到视觉、听觉、触觉等多个交互通道,彼此能力差异、数据互斥,同时又不可避免地产生相互干扰、遮蔽等问题,为使各种交互通道达到最佳协同模式,使人获得自然交互感受,必须建立智能化的新机制。

2　虚拟(增强)现实微显示关键技术

2.1　基于硅基的微型OLED显示技术

硅基OLED显示技术,是结合CMOS工艺和OLED技术的一个综合技术,它使用单晶硅作为驱动板,在其上直接制作而成[5]。不像LCD或OLED将

屏幕建立在玻璃基板上,硅基OLED直接搭载在晶圆,像素尺寸在4~20 μm,而标准OLED面板的像素尺寸为40~300 μm。硅基OLED主要是在两层电极之间使用能够发光的萤光有机材料,电流通过后会发出单色光,再透过滤色器生成所需的颜色。

硅基OLED使用单晶硅为基底,在基底上使用CMOS工艺制作电路,作为OLED所需的驱动和数模转化等电路。由于硅基不透明,还需要将OLED的光发射出去。首先在单晶硅基底上用高反光材料制作阳极,使其具有较高的发光亮度。然后制作OLED发光的主体单元,包括空穴注入层、空穴传输层、发光层、电子注入层、电子传输层等有机半导体层。最后为了光从顶部发出,使用半透明金属作为阴极,并在阴极上制作透明封装层进行杂质保护,并且进一步贴合玻璃进行强度保护[6]。

硅基OLED由于不需要照明光学系统,通常比硅基液晶(LCOS)更紧凑。硅基OLED就是在维持相近分辨率水平的基础上显示面积更小的OLED,这一特性使它拥有了更高的像素密度,且具有自发光、厚度薄、质量轻、视角大、发光效率高、对比度高达10000∶1等特性,广泛地运用在各类近眼显示中。硅基OLED具有更快的响应速度,色彩效果更丰富、可实现高分辨率,采用全固态器件,具有工作温度范围宽、抗震性好、集成度高等优异特性,可用于数码相机、电子取景器、无人机FPV和AR等领域[7]。

目前,硅基OLED技术可实现3000~5000 cd/m^{2}的显示亮度和每英寸3000像素的分辨率。但是对于AR系统,其亮度还远远不够。且显示屏寿命相对较短,色彩纯度不够等因素,在AR显示方面的研究还需要进一步探索,关键要提高微型OLED显示器的显示亮度,器件寿命,以及高可靠器件的规模量产[8]。

2.2 MicroLED显示技术

MicroLED是一种全新的显示技术。在芯片上直接集成高密度的LED阵列,每个像素都能独立驱动,具有高分辨率、长寿命、效率高、色彩饱和度高、高动态范围等优点[9],被认为是VR/AR微显示的最佳解决方案。其刷新率、亮度、发光方式、像素密度等指标经行业内多家机构测试均具有领先优势。

MicroLED技术即LED微缩化和矩阵化技术,是指在芯片上直接集成高密度的LED阵列,是将LED进行矩阵化、微缩化和薄膜化的结果。MicroLED技术可以将目前的LED微缩至长度仅50 μm左右,是原本LED的1%,比一粒沙还细小。MicroLED不仅限于背光源,其可直接将R、G、B三原色的芯片拼成一个像素点,变成“1个像素”的概念,不再需要滤光片和液晶层。每个MicroLED都是一个像素,并且可以单独点亮,这样的改变使整个显示模组更加精细,其显示亮度、画质、反应速度都有大的提升。MicroLED典型结构是一个半导体器件,由直接能隙半导体材料构成。其工艺路线为:获取共阴极MicroLED单片阵列,根据MicroLED单片阵列完成COMS像素电路以及驱动背板的设计与制备,在MicroLED单片阵列及驱动背板上制备高精度互连的键合材料,然后在具有高精度对位的键合设备里完成MicroLED与驱动背板的键合,接着在上述组合结构上进行光效提升、防止光学串扰及色转换层的制备。

目前,MicroLED尚未产业化,且面临许多技术挑战,亟待解决的技术瓶颈包括外延晶圆工艺、巨量转移、背板与驱动技术及全彩色显示技术等。其中有很多困难需要解决,从前期的磊晶技术瓶颈、巨量转移良率、封装测试问题,到后续的检测、维修等都是很大的挑战。在显示效果方面,MicroLED具有高解析度、低功耗、高亮度、高分辨率、高对比度、高色彩饱和度、反应速度快、厚度薄、使用寿命长等优势,同时具有发光效率低、热损耗高等劣势。随着海兹定律(指每18~24个月LED的成本价格将降低1/10)推动LED成本持续下探,芯片尺寸不断下降,MicroLED量产成为可能,届时MicroLED可应用在AR/VR、智能手机、平板电脑、高阶电视和可穿戴设备等多个领域。

3 虚拟(增强)现实光学系统关键技术

3.1 表面浮雕光栅波导技术

表面浮雕光栅波导以其设计自由度高、衍射效率高、可复制性好的优点,成为当前AR衍射波导光波导的主流方案之一。传统的折反射光学器件

(refractive optical element,ROE)对于光束调制能力有限,表面浮雕光栅(surface relief grating,SRG)属亚波长尺度元件,表面浮雕光栅光波导使用SRG作为光波导系统耦入、耦出和出瞳扩展器件,相较传统方式对光束的调制能力显著提高[10]。SRG通常指的是光学表面布设的各种不同周期性、几何参数的凹槽结构。SRG在设计加工时主要通过设计凹槽的轮廓、形状和倾角等结构参数调整衍射效率和成像质量,常用的SRG结构分为一维光栅和二维光栅,前者典型如矩形光栅、倾斜光栅和闪耀光栅等,后者典型有柱状光栅。SRG光栅厚度属微纳米级别,可以直接在表面制备,这很大程度上减小了AR系统的体积。随着半导体制备工艺的进步,SRG光栅结构也在不断优化,加工手段日趋成熟,市面上大多数AR产品使用了SRG技术[11]。

微软的HoloLens系列以及WaveOptics公司的系列产品是表面浮雕光栅波导最具代表性的应用,消费级产品的问世证明了它的可量产性。表面浮雕光栅波导的设计依赖于复杂的严格耦合波(RCWA)理论,制造依赖于微纳制造工艺,在微纳尺度的设计难度依然较高,设计与制造时的微小偏差都会影响产品质量,对加工的高精度要求也导致成本提高。当前已经被设计并制造出的产品在具体成像和提高视场方面难以兼备,产品的用户体验难以保证,因此表面浮雕光栅波导的技术方案和制造工艺仍是目前AR终端制造最核心的问题之一。

目前,表面浮雕光栅波导技术的研究重点是研究表面浮雕光栅波导的优化设计,攻克纳米压印模板制备和压印工艺问题,实现基于纳米压印方法的表面浮雕光栅波导的量产制备。

3.2　体全息光波导技术

体全息光波导与表面浮雕光栅技术类似,不同点在于体全息光波导中的耦入耦出衍射光学元件为体全息光栅(VHG)。VHG拥有非常好的光学特性和设计灵活性、优越的稳定性和一致性[8]。VHG特性衍射主要体现为布拉格体效应,当入射光满足布拉格条件时,VHG的衍射效率就能够保持在较高等级,而如果达不到布拉格条件,VHG衍射效率就会显著下降。因此VHG厚度方向上的调制为光栅引入了一定的选择性,继而可获得某一特定的衍射级次。为获得折射率周期性变化VHG,VHG加工常用双光束全息曝光技术,干涉结果记录在具有一定厚度的感光材料内形成干涉条纹。

自2000年美能达公司首次提出将VHG应用到衍射光波导的体全息光波导方案,VHG就成为光波导技术研究重点之一。随着技术的不断成熟,各家公司陆续推出自己的体全息光波导方案。DigiLens公司提出的方案是以可切换布拉格光栅(SBG)为主的衍射光波导,并且针对布拉格光栅的特点进行优化,设计了一种可切换布拉格光栅阵列结构,该结构依靠多个布拉格光栅混合使用实现彩色成像,同时该结构有效地减少了系统颜色串扰。索尼公司出过一款高亮度的单色体全息光栅波导,该方案构建一个采用双面体全息光栅作为出入耦合端的系统,系统性能上能够达到85%的透射率。此外,英国的TruLife和WaveOptics,以及美国的Akonia公司也提出了各自的体全息方案。总体而言,目前VHG方案主要通过优化每个光栅可切换的参数,使光栅在特定范围内调整光束,但目前这些公司的体全息光波导方案的效率不高。

VHG结构可有效减小系统的占用体积和实际重量,容易实现集成其他光学器件功能,对于系统微型化和轻薄化设计作用巨大。在成像上,VHG元件还具有较好的色彩均匀性,在实现单片彩色波导方面也有独特优势。VHG的加工难点在于感光材料的选择,对于材料噪声颗粒、环境适应性都有较高要求,难以实现大规模的量产。同时VHG的响应带宽较窄,会带来视场角较小等问题,在使用VHG做大视场角系统时一般需要叠加多层全息光栅,对于多层结构的工艺技术要求进一步提高[12]。目前来看,体全息设计方案的进一步成熟和量产良率的提升预计还需要一定的时间。

体全息光波导技术未来在于研究全息光波导的3D光刻技术,搭建微纳尺度全息光波导3D光刻系统,利用双光子聚合原理,突破衍射极限,实现对模拟设计的全息光波导光栅进行加工制备。开展基于几何阵列光波导的紧凑型全彩显示系统研究,实现符合人体功效学的大视场角彩色双目波导显示系统。

3.3 超表面光波导技术

超表面是一种人工制造的、通过在光学表面上的微结构单元来实现光场调控的新型光学元件[13]。超表面的光场调控研究工作起源于人们对斯涅耳定律,即光的折射定律的重新探索发现。根据斯涅尔定律,传统光学元件通过光波在不同几何形状和折射率的介质内传播的过程中积累相位,从而改变光波波前,实现光场调控。但是,天然光学材料的折射率限制现有的传统光学透镜尺寸进一步缩小。Yu等[14]2011年提出了广义斯涅耳定律,并设计出了新型光学元件超表面。

从超表面研发出来至今,因其具有创新的物理机制、灵活的结构设计等吸引了一大批研究者进行开发和研究。超表面可灵活地选择不同调控机理的单元结构实现所需相位分布,从而实现超透镜成像、消色差超透镜阵列集成成像、超表面全息图成像等多种功能,在AR近眼显示领域展现出巨大的应用前景[15]。Li等[16]2021年设计和加工出大尺寸厘米级别的消色差超透镜,解决了消色差超透镜尺寸受限问题;Lee等[17]研究设计出超表面全息图,轻薄的超表面全息图有望应用于AR近眼显示系统的图像源实现三维显示。西安电子科技大学王晓蕊团队在近眼显示及基于超表面的集成成像三维显示方面开展了研究工作,并于2018年设计出多种提升集成成像显示质量的超表面结构[18];南京大学、中山大学和四川大学王琼华等团队则使用消色差超透镜阵列实现集成成像光场显示[19]。

超表面的出现为解决传统AR近眼显示系统设计中的受限问题以及实现光学系统集成化和微型化提供了有效的技术途径。但是想要利用超表面实现大视场、消色差、轻薄紧凑结构和高分辨率的AR近眼显示系统,仍需要克服理论和实验方面的多重挑战。

4 虚拟(增强)现实感知交互关键技术

4.1 手势识别交互技术

4.1.1 手势识别的关键技术

人机交互通俗来讲就是人与系统之间的沟通,互相传达信息,而这中间传达的信息正是简单的指令。人们将简单的指令传达给系统,系统来完成所获得的任务,之后再给予人答复。这里所说的系统可以是计算机的系统和软件,也可以是各种样式的机器。人机交互研究的技术也有很多,并且一直以来都是炙手可热的研究应用热点领域。这些热点领域有手势识别技术、语音识别技术、眼动追踪交互技术、压力触控技术等。

在众多人机交互技术中,手势是人与人以及人与机器交互最常见的通信方式之一。手势相比语言更加简洁,常常会忽略很多情绪信息,使得传达的信息内容更加简明扼要。由于手势的快速、简单和自然等相关特性,使得手势被运用到了人机交互过程的很多领域。人们可以使用手势识别方式,更自然、更高效地和电脑实现互动。此外,由于手势识别技术也被广泛运用在电脑游戏、手语互动、文字阅读系统、智能家居等领域,所以手势识别也成为人类科学研究的热门技术之一。

根据不同设备输入模式的不同,手势识别可以划分为基于数据手套的手势识别和基于视觉的手势识别,前者因为涉及到数据手套的购买,所以需要大量费用。因此,近年来学术界以及商业界所研究最多的是基于视觉的手势识别技术。而基于视觉的手势识别又可分为静态手势识别和动态手势识别2种类型。在静态手势识别当中,要根据图像中的信息,判断出手势的类别。而动态手势识别并不单纯依靠单一图片来判断手势,还需要根据视频中的时间信息来判断手势,即利用时空信息来跟踪整个手部。因此,手部检测以及手部跟踪往往在手部动作分类之前完成,作为动态手势识别系统的第一个步骤。动态手势识别相较于静态手势识别的优点很明显,静态手势识别只能根据单张图片来判断手势,但是人们的手势往往是以一连串信息表示的,只有少数手势是一个单独动作就能表达出来的。因此,可以根据时序信息判断动作的动态手势识别技术脱颖而出。在现代人工智能时代,很多方面都应用到了动态手势识别技术,例如有利于聋哑人与大部分不会手语的正常人之间的手语识别技术,方便司机与驾驶车辆之间进行沟通的智能驾驶技术,利用一连

串手势就能控制家具开关的智能家居技术等。由此可见,动态手势识别已经成为人们日常生活、工作和学习中不可缺少的一项重要技术。

4.1.2　手部检测研究现状及难点

为了去除环境中的干扰,深挖出更加有用的手势信息,往往需要引入手部检测,这也是整个手势识别流程中比较重要的步骤。手部检测的方法有很多种,最原始的方法是人为提取图片上手部信息的特征[20-22],这种手部检测的方法也可以分为基于手部形状、肤色[23]、运动信息[24]等方法。这些传统方法都有一系列通病,例如受光照、自遮挡、手部形状以及肤色差异等众多因素影响而导致实验结果产生较大差异;并且这类方法的参数量较大,计算速度较为缓慢,很难达到实验所要求的实时性。现如今使用较多的方式是深度学习。

手部特征可以利用深度学习的方法[23-25]进行提取。深度学习的方法针对手部特征的提取有更强的鲁棒性。Zimmermann等[23]使用HandSegNet方式进行手部特征的提取。其中,HandSegNet卷积神经网络一共有16层,它的工作方式是输入光学三原色片,返回的图片是2通道形式的;返回的图片由手部图片和背景图片共同组成。通过将手部图片以及原图片进行组合,就可以轻松得到手部区域。随着深度学习的不断发展,基于候选区域的手部检测方式也应运而生。Grill等[26]使用调整后的Faster R-CNN网络检测每一帧图片中手部的位置,消除场景对手部的影响。Liu等[27]通过融合了的RGB和Depth深度双通道信号获取手信号,手部检测的准确性有了提高,同时鲁棒性也获得提升。这种方式主要是在原有的Faster R-CNN网络中融入了手的深度信号,并与RGB信号一同实现了特征融合等一系列运算,随后再使用区域候选网络RPN获得可能是手的若干候选框,通过对可能有用的手区域进行池化、分类以及回归等运算,最后得到比较精确的结果。

除了以上几种方法,为了解决大幅度的手部动作以及双手动作带来的一系列问题,Narayana等[28]利用two stream Faster R-CNN以及获取手部骨架信息的方式,先进行手部范围的检测,再利用提取出的手部骨架信息进行左右手的区分,为手部检测的发展提供了良好的参考方向。这种方法还解决了手部检测的一个难题,那便是手部在整张图像中的占比是不可预知的。他们利用全局信息以及局部信息相结合来表达整个手部的运动过程。其中,全局信息指的是整个视频的序列帧中所包含的信息,例如整个手部运动的一系列细节、背景变化、手臂等非手部肢体运动信息等。局部信息指的是手部运动的整个过程,手部检测器首先检测出视频中每一帧的手部位置,检测出手部位置的每一帧图像又组成了局部信息。当识别大幅度手部动作,或是人体其他部位信息以及背景信息对于手部动作有相关提示的情况下,全局信息可以是有利的工具,但针对一些细微的手部动作,例如勾手指等,单独使用全局信息往往达不到很好的检测结果。因此,高准确性以及高鲁棒性的结果要归功于这种将全局信息以及局部信息相结合的做法。

2015年YOLOv1横空出世,引起了深度学习领域人员的高度重视。YOLO的全称为"You only look once",在名字上就体现了YOLO算法运算迅速的特点。与R-CNN不同,YOLO算法可以巧妙地归结为深度学习中的回归问题。YOLOv1的核心内容是将需要的图片作为输入,在输出端回归处理Bounding Box位置以及类别。2016年YOLOv2被提出,运行速度以及识别精确度都比YOLOv1有所提高;除此之外,它可识别的对象提升到了9000种;该方法称为联合训练算法。2018年作为YOLOv2的升级版的YOLOv3网络模型被提出,YOLOv3实用程度远超YOLOv1和YOLOv2,并且网络模型也更加复杂。YOLOv3引入了被大家关注的多尺度预测,这也使该网络在许多目标检测中得到了很好的应用,例如监控下的人员异常行为监测。另外,YOLOv3还使用了darknet-53作为分类网络,取得了较大成功。2020年YOLO算法的提出者Redmon[29-31]退出CV届,2020年4月,曾经在Redmon团队的Bochkovskiy及其团队[32]公布了YOLOv4,YOLOv4可以称作是一个结合体,结合了SPP、CSP等结构,取得了很好的实验结果。在2020年6月YOLOv5被提出,YOLOv5的作者并没有将

YOLOv4与YOLOv5进行比较,但是YOLOv5的模型更加小巧,仅有27 MB,远小于YOLOv4,而识别准确度也与YOLOv4不相上下。YOLOv5共有YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x这4个模型结构的深度和宽度都不一样,YOLOv5的出现使得许多人不再使用原来的v3版本,而是直接使用YOLOv5进行目标检测,例如吸烟监测、安全设备监测以及异常行为监测等。值得一提的是,与以往的YOLO系列相比,YOLOv4以及YOLOv5都使用了mosaic增强,这对于小物体的检测识别是很友好的。

4.1.3 手部建模与动作分类研究现状及难点

动态手势识别的核心是动态手势的辨别,即如何利用输入的信息有效的识别出手势。显而易见,动态手势的辨别与动态手势建模以及分类网络是密切相关的。将近年来的识别方法进行总结,可以大致分为以下2种分类:基于机器学习的方法[33]以及基于深度学习的方法[34-35],其中前者是基于人工特征的,后者是基于自主特征的。

基于机器学习的方法在2019年之前十分流行,其中动态时间规整算法(dynamic time wraping,DTW)[36-37]以及隐马尔可夫模型(hidden markov model,HMM)[38-39]是其中比较出名的方法。

动态时间规整算法最早是被应用在语音识别中,用来检测不同长度序列相似程度的一种方法。Tang等[40]在2018年发表了将DTW应用在动态手势识别中,获得了不错的识别效果。利用动态时间规整算法进行手势识别的工作原理在于——先预处理手势,将这些训练集中的每一帧提取特征并进行归一化形成模板,经过同样步骤的测试手势所输出的结果与之前训练的序列模板进行匹配,距离最小的即为认定的手势,算法实现流程图如图1所示。但是当待处理的数据量较大或者手势相对复杂时,没有使用统计模型训练的DTW算法就不能够很好地被使用,因此张建荣[41]将动态时间规整算法进行修改,并于2016年发表了改进的动态时间规整(Improve Dynamic Time Warping,iDTW)进行动态手势识别。iDTW的工作原理是提出点以及面组成的范围来约束路径,根据各节点的距离方差,来进行动态地分配各节点上的权值。对比DTW与iDTW,后者不论在准确度还是运算量上都更加优秀。

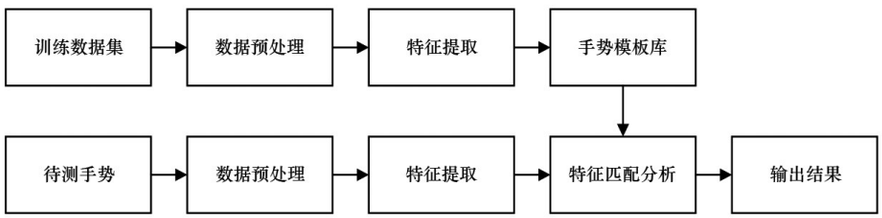

图1 动态时间规整手势识别算法流程

隐马尔可夫模型是一个使用较多的经典模型,它可以处理时间以及状态等序列问题。在HMM模型中,有2种类型的数据是需要使用到的,一类是观测序列,另一类是隐藏数据。其中在动态手势识别任务中,同样总结出这2类数据——状态序列和隐藏序列。状态序列即是用手实际去做的各种各样的动作,而隐藏序列则指的是用手部发出的动作。基于隐马尔科夫模型的动态手势识别算法,是要在观测序列中挖掘出隐藏数据所涉及的内容,用HMM模型将各类型手势一一对应。在练习过程中,首先必须要将训练集的样本类型一一区分开来,而后再通过先向和后向的计算,将各个类型的手势训练出属于该类型的HMM模型。在测试过程中,通过先向运算,将待测数据集与所有HMM模型进行了匹配分析,得到每个HMM模型的概率数值,一个HMM模型与每类手势一一对应。训练过程中,需要先将训练集样本类别一一划分出来,然后根据前向以及后向的算法,将不同类别的手势训练出属于这一类别的HMM模型。测试过程中,采用前向算法将待测数据集与所有HMM模型进行匹配

分析,以此获取每个HMM模型的概率值,概率值最大的结果即为最终输出结果。Saha等[42]利用HMM算法进行动态手势实验,实验数据为60类不同主题背景下的12种不一样的动态手势,每一类的正确率都在90%左右。

基于机器学习的方法与基于深度学习的方法相比,基于机器学习的方法很难在不同角度和不同环境下,设计出一种所有实验数据都适用的特征提取方法,因此基于自主学习特征的深度学习方法应运而生并得到各界重视。近年来,随着深度学习时代的到来,深度学习也被应用到动态手势识别领域。利用深度学习的方法进行手势识别的核心关键内容是利用搭建好的神经网络进行后续一系列操作,即首先初始化所搭建的模型,之后进行前向传播等操作,所搭建的网络利用手势的数据不断学习,自身不断优化。数据集的质量与神经网络训练得出的结果息息相关,深度学习的重中之重就是需要搭建出优秀且易于使用的神经网络。现如今,用于动态手势识别操作比较出名的深度学习方法有双流法(Two Stream)[43-45]、3DCNN[46]等。动态手势识别建模经典方法汇总,如图2所示。

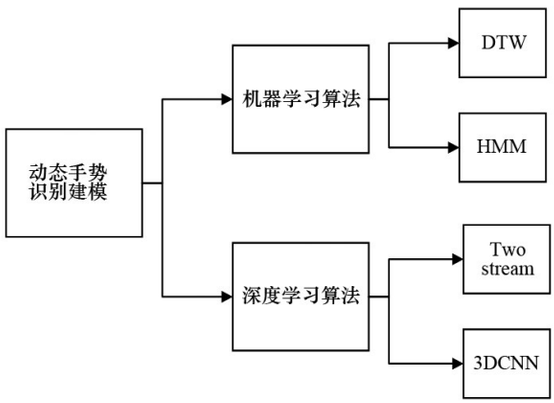

图2　动态手势建模经典网络汇总

双流法的网络中包含2个网络结构,分别是空间网络(spatial network)和时序网络(temporal network),前者负责的是提取图像中所表现出动作的空间信息,后者负责处理的则是帧与帧之间手部动作的光流信息。空间与时间网络往往是应用ResNet、AlexNet以及VGG等经典网络进行识别的。双流结构是在2014年由Han团队[47]提出的,该网络的空间与时间网络采用的是VGG。该算法的空间网络输入与时间网络输入分别是RGB图像以及堆叠的5帧光流图像。这样做所造成的不好结果就是该网络只能进行短期的行为识别,针对那些持续时间较长的动作并不能很好地识别。2016年时域分割网络(temporal segment network,TSN)的诞生解决了这一问题,它是由Feichtenhofer等[48]提出的。本算法的主要改进之处在于采用了稀疏采样的方式以识别长视频的动作,除此之外,还采用了Inception v2[49]进行空间网络以及时间网络的构建,增强了TSN算法的鲁棒性。同年,Lin等[50]又在时间与空间信息的信息融合上做了改进,使得整体网络的参数量有所下降。国内外大量学者利用two stream算法进行手势识别以及行为识别,经过实践证明two stream是视频级行为识别的优秀方法之一,但是它也存在一些弊端,例如参数量大以及光流图提取耗时且难度大等问题。

另一种经典方法是3DCNN算法。3DCNN算法可以直接在序列中提取出每一帧的空间信息以及帧与帧之间的时序信息。3DCNN算法被行为识别届的学者广泛应用并且衍生出了很多的变体,其中最经典的方法是2018年由Yan等[51]提出的。8个卷积层、5个池化层、2个全连接层和最后的softmax分类层共同构成了它的网络结构。动态手势识别领域中,许多学者也尝试使用了可以高效提取时空信息的3DCNN网络,Zhou团队[52]在2020年将3DCNN网络应用到了动态手势识别中。Zhu等[53]在网络中采用金字塔输入对手势进行划分,采用金字塔融合进行特征融合。Li等[54]结合3DCNN巧妙地从深度以及梯度图像中识别出了驾驶员的手势动作。由于3DCNN网络结构简单,导致最后特征的提取以及表达上被制约住,但是加深网络结构的话又会使原本参数量不少网络的参数量更多。为此Tran等[55]在2017年提出了ResC3D网络,将残差块加入到3DCNN中,利用恒等函数来加深网络。之后Miao等[56]在手势识别的相关任务中应用了ResC3D网络,并取得了不错的效果。并且Li等[57]在该网络中加入了注意力机制,使得特征提取效果

提升。3DCNN的弊端在于它的数据量巨大,在进行训练时使用的时间往往会很长,如果再与其他网络进行结合会造成在预测过程中有较大的延迟问题。

手势识别与行为识别有一个比较大的区别就是手势识别对于时序信息的要求更加严格,因此处理序列数据的循环神经网络(recurrent neural networks,RNN)被应用到了手势识别领域中。现在应用较多的网络是RNN的一个变体——长短期记忆模型循环神经网络(long short term memory,LSTM)。LSTM很好地解决了RNN在训练过程中梯度消失及爆炸等问题,在长视频上有更好的应用效果。许多学者在近几年将很大的精力集中到了循环神经网络与3DCNN网络相结合的应用中,例如Gupta等[58]将3DCNN与RNN相结合提出R3DCNN,该算法的主要思路是利用3DCNN网络来处理较短的时序信息,之后利用RNN处理较长的时序信息。将RNN与3DCNN网络进行级联,增强视频中前后时序信息的关联。由于RNN的弊端,Zhu等[59]将3DCNN网络以及LSTM进行级联,成功应用到了手势识别领域。LSTM也有很多变体,门循环单元(gated recurrent unit,GRU)就是其中之一。LSTM和GRU在网络性能方面是水平相当的,参数较少的GRU比LSTM更易得到收敛,因此针对数据集较少的实验,GRU表现出来的效果往往要强于LSTM。

不论是在学术界还是商业界,动态手势识别作为学科热门,以上分析中所有方法性能的优劣不能一概而论,需要根据针对不同场景下的不同要求,选取最合适的动态手势识别的方法及其网络。

4.2 脑电识别交互技术

脑电是中枢神经系统产生的一系列非平稳变化的空间离散随机信号,通过头皮记录的电位变化表现出来,相比其他生理信号,脑电信号(electroencephalography,EEG)对人的情绪状态有更加真实可靠的反映[60]。

19世纪80年代开始,脑电信号的探索逐渐展开,人类历史上第一次采集动物生理活动的脑电信号发生在英国,外科医生卡尔顿通过电极检测器来检测猴脑的电生理活动。之后到20世纪,随着科学的发展,欧洲医学界首次使用像针一样的电极记录人脑电流信号变化状态,开创了脑电图记录方案[61],标志着脑电信号在临床阶段应用的开始。20世纪30年代,随着模拟电子技术的发展,使用电子管技术开发的脑电信号放大器采集系统发展起来。自20世纪50年代以来,计算机技术作为一种新的计算技术开始迅猛发展,并被应用于脑电图的研究中。

1936年在医疗诊断中开始初步应用脑电信号分析系统[62]。1991年,Gabor等[63]提出将“特征提取”与“监督分类器”两者相结合的设计思想,设计出用于检测癫痫病的人工神经网络并首次在临床检测中得到应用。其所使用的设计思想成为探究脑电信号工作机制和工作原理的主流设计思路。如何设计更好的分类器和特征提取方法已成为脑电分类研究的热点。目前卷积网络的设计、特征数据挖掘方式、信号分析算法的研究占据重要的地位[64]。国内外主流的脑电信号分析方向主要是时域、非线性动力学[65]、频域[66]、时频域4个方向。根据历年脑科学大会研究进展交流,在脑电信号预测方面目前国内外主流的方法大致基于以下几种算法:模拟线性预测、非线性逻辑回归预测、基于径向基函数以及叠加分类树判别效果的神经网络、离散小波变换[67]、经验模态分解[68]中的本征模(empirical mode decomposition, EMD)等技术,其中癫痫脑电图诊断当属于效果最成功的临床医学案例[69]。基于傅立叶变换和决策树的混合框架是比较成功的分类器,假设脑电信号仍然是一个传统的电流信号,采用傅氏变换用于特征提取,模式识别分类算法则采用决策树算法。典型分类器则采用希尔伯特黄变换(Hilbert-Huang Transform,HHT)特征挖掘结合支持向量机(support vector machine,SVM)模式识别算法[70],对癫痫发作时的脑电信号能量谱进行判别分类。根据传统图像分类,从脑电信号灰度图像的直方图中,挖掘出像素强度属性,进行多维特征融合,融合后模型分类准确度可达到99.125%。20世纪70年代以来,采用集成运算电路技术和共模信号放大技术[71],采集脑电信号的硬件系统变得更加便捷。同时,脑电采集系统的性能也

大大提高。从此,脑电采集的硬件设备逐步完备,为脑电信号研究提供了强有力的支持,使得脑电信号的探索研究发展进入到新阶段[72]。

人脑产生的脑电信号特征隐含人的真实意图,科研人员能否从没有规律的脑电信号中挖掘出相似的人类意图,以及设计出最佳模式识别的系统是脑电信号分类系统的关键,也是重要的因素。尽管所采用的技术方案与分析理论对于国内外研究团队均有所不同,但是如何把参与者的思想意图完美地表达出来,是所有脑科学研究者所面临的"卡脖子"问题,也是将脑电信号用于辅助诊疗实践、实现各国脑科学战略目标的唯一途径。

越来越多的研究也证明了脑电信号、心理状态两者之间存在某种特定的相关性,例如可以通过脑电信号心理分析做出相关决策,并且在康复诊疗方面得到运用等。因此,脑电信号信息挖掘与生理信号模式识别研究技术也被21世纪的科学家认为是一种具有重大历史意义的技术前景。与此同时,随着人工智能技术的发展,深度学习在诸多方面得到广泛应用。深度学习以强大的学习能力以及针对数据流的非线性拟合能力,逐渐开始被应用到脑电信号特征挖掘、脑电信号模式识别任务研究中,并且取得了非常好的效果。

在神经学科方面,研究脑电信号的规律以及分析工作的主要目的是对照大脑的工作方式以及运行机制,辅助诊断某种特定神经类型的疾病。例如,癫痫患者大脑功能区发生异常障碍时,患者头部神经元细胞突然异常放电,此时可根据脑部功率变化的规律对癫痫病患者进行识别。目前,主流癫痫病的诊断方法是神经科专家根据观察记录的脑电信号是否存在与正常脑电信号不相同的异常脑电波。但是,由于癫痫患者发病时间没有任何规律,为了防止癫痫患者发病影响自己日常的生活状态,因此设计出病发预警系统。该系统首先能够满足24 h监测到癫痫病症状的出现并预警,同时要求具备良好的续航能力,预警系统准确度要高,检测用时要少等,达到以上几点就可以进行推广运用。因此,使用AI+脑电的实时监测系统在技术上、实践上都是可行的,在降低疾病对人本身造成伤害层面具有重要的意义与价值。除此之外,相关的神经系统疾病均可以采用脑电信号作为依据,实时预测发病时间、状态。

在心理学上,脑电信号也非常有用。例如,利用脑电信号调查消费者的消费者心理。如果消费者确认了不同的价格,可以根据脑电信号的态势变化来定制价格,使产品更加符合消费者的心理。另一个例子是使用脑电信号分析犯罪者的精神活动。由此,当犯罪者被问讯时,可以根据犯罪者脑电信号的变化来推测犯罪者的心理活动,制定更准确的问讯计划。

从认知科学的观点来看,使用脑电信号可以检测司机是否疲劳、注意力是否集中等驾驶状态,根据驾驶员的状态来提示司机,还可以使用脑电信号识别患者的情绪状态,对患者进行更适当的治疗,提高治疗效果[74]。尤其针对某些特定人群无法表达自己的真实感受,或者内向不愿意表达自己的真实想法,但是在治疗过程中医生又必须去了解患者的倾斜屏的状态,选择使用这种方式无疑是更加科学,更加适合的。此外,学生可以通过脑电信号实时获取自己的学习状态,如注意力是否集中、心态是否良好等,根据这些学习状态特征针对性的调整学习战略,提高学习效率。

在医学上,使用脑机接口技术,使失去部分运动能力的患者能够通过脑电信号意念[75]控制轮椅的状态——行走、停止、转弯和后退,为生活提供了便利。这项研究的关键在于对脑电信号的识别。目前,有很多研究机构从事脑机接口(Brain ComputerInterface, BCI)的研究[76],国外主要有Wadsworth研究中心、Graz的BCI研究小组和N. Birbaumer实验室,国内主要有清华大学、华南理工大学和香港科技大学等。目前,主要采用MI、SCP、SSVEP、Mu节律、ERD/ERS和P300等典型单一模态的脑电信号来实现脑电采集,其中清华大学、天津大学、北京邮电大学在2020年BCI脑意念控机器人大赛中获得优异的成绩。

基于脑电信号在生物反馈信息挖掘方面相关研究。1984年,Sutter等[77]已经通过基于瞬态VEP信息研究出大脑反应接口的实时BCI系统。用户

可以在显示屏上查看8×8红色/绿色待机不同符号样子矩阵,从所有的符号样子里选择目标的符号。实验结果表明,受试者平均每分钟可以打出10~12个给出对比的英文单词。伊利诺伊大学的Farewell和Donchin根据事件相关电位的成分,由刺激诱发的潜伏期约300 ms晚期正波的P300设计了一种基于脑电波幅与时间相关潜伏的虚拟打字机[78]。该虚拟打字机实际通信速度可以达到每分钟7~8个字符左右。根据脑电的ERD/ERS特征,Pfurtscheller研发了Graz-BCI系统[79]。被试者在这个系统中经过若干天训练,数据分类准确率达到85%以上。然后,再基于运动想象脑电的ERD/ERS建立了基于左右脚或舌的3种类型的异步运动图像BCI系统。这3种类型的分类准确率均可达到83%~85%,平均单个单词的拼写所占用的时间约在0.4~1.0 min。目前,针对脑电信号Mu节律的研究探索中,Wolpaw等[80]依据受过训练的受试者可以改变Mu节律的幅度的特性来控制光标的移动。设计的系统经过6~8周左右的训练,其中一部分受试者的正确率达到70%。但是实验结果同时表现,并非所有受试者都可以从中挖掘到这种生物反馈控制特征,准确率也是因人而异。

目前,脑电识别研究取得了一系列积极进展,部分研究也逐渐走向应用。但是,脑电识别还有很多方面值得研究,较多关键性的问题也有待进一步解决。主要有以下几个方面:

1)脑电识别模型需要数据集进行训练,如针对某一种情感进行识别就需要相关情感的脑电数据,对于没有相关脑电数据的情感通常无法识别。

2)脑电识别模型的泛化能力较差,由于识别率与研究对象有关,脑电数据太少会导致过拟合的问题。

3)脑电信号极易受到噪声干扰。脑电信号幅值较低,极易受到电磁污染;脑电采集设备传感器容易受到温度影响,发生直流偏移物理现象;被试自身的生理反应——眨眼、呼吸、脉搏、心跳等干扰的存在,导致采集到的脑电信号存在大量噪声污染。

目前研究趋势为:

1)为提高模型的准确率与泛化能力,脑电数据集向更大更多元的方向发展,识别模型也由简单模型向复杂的混合模型发展,多种网络模型结合在更大的数据集上训练从而提高模型性能。

2)多学科交叉融合进行特征设计,将神经科学与心理学的最新研究与脑电识别相结合,多种生理信号和特征结合使用,多学科交叉来寻找与情感相关的信号特征。

近数10年来,由于脑电信号分析和模式识别技术快速发展,基于脑电信号的应用开始在若干个领域出现。但是传统的机器学习算法在脑电信号分类方面运用遇到了很大障碍——轨迹干扰噪声的存在,导致在各种领域脑电信号的广泛应用受到了限制。与此同时,深度学习技术迅速发展,取得了惊人的成果。考虑到深度学习超强的自动特征提取功能和非线性拟合功能,深度学习方法有望打破以往机器学习的瓶颈。这也表明,基于神经网络算法的脑电信号分析和模式识别技术的研究,对脑电信号在各个领域的应用非常重要。随着脑电识别技术的不断进步,脑电识别技术研究会取得更大进展,不断优化基于脑电信号的情绪识别方法,并在实际产品中得到更好的应用,发挥出脑电识别研究的应用价值。

5 结论

综述了元宇宙终端——虚拟(增强)现实技术在微显示关键技术、光学系统关键技术、感知交互关键技术的发展现状。重点介绍了硅基OLED微型显示技术、MicroLED显示技术、表面浮雕光栅波导技术、体全息光波导技术、超表面光波导技术、手势识别交互的关键技术、脑电识别交互技术在虚拟(增强)现实中的应用。

目前,硅基OLED微型显示技术占主导地位,但是随着MicroLED微显示技术在量子点、巨量转移等方面的突破,未来MicroLED微显示技术会有更好的发展前景。由于超表面光波导非常轻薄,天然适合在虚拟现实中应用,在其数值孔径等关键技术指标取得突破后,将被广泛应用。人机自然交互,特别是脑电识别交互技术是另一个研究重点,

也是制约元宇宙发展的关键,需要准确捕捉人体语音、动作、眼球等各种信息,实现准确操控,需要继续提升视觉、听觉的模拟效果,并在触觉、嗅觉、重力感觉等方面取得突破,从而实现自然交互。

参考文献(References)

[1] 崔迪. 面向建筑信息的多人虚拟交互方式研究——以六主村无止桥公益项目情景为例[D]. 上海: 同济大学, 2018: 1-3.

[2] 赵沁平. 从虚拟现实技术管窥新兴工科人才培养[J]. 中国大学教学, 2019(9): 7-9.

[3] 史晓刚, 薛正辉, 李会会, 等. 增强现实显示技术综述[J]. 中国光学, 2021, 14(5): 1146-1161.

[4] 王伟. 光波导成AR眼镜新宠[N]. 中国电子报, 2021-11-23(005).

[5] 王伟. 硅基OLED微型显示领域又有新进展[N]. 中国电子报, 2022-06-28(06).

[6] 杨建兵, 秦昌兵, 张白雪, 等. 大尺寸高分辨率硅基OLED微显示技术研究[J]. 光电子技术, 2019, 39(3): 181-185.

[7] 张天宇. 京东方显示屏出货实现全球"双冠"[N]. 北京商报, 2019-11-22(F5).

[8] 史晓刚, 薛正辉, 李会会, 等. 增强现实显示技术综述[J]. 中国光学, 2021, 14(5): 1146-1161.

[9] 陈浩, 朱杰辉, 沈庆云. 一种舞台用快装式LED灯装置: 202210900390.2[P]. 2022-07-26.

[10] 姜玉婷, 张毅, 胡跃强, 等. 增强现实近眼显示设备中光波导元件的研究进展[J]. 光学精密工程, 2021, 29(1): 28-44.

[11] Richter P, Bürger A, Waldern J, et al. Compact AR-HUD solution with optical waveguide[J]. ATZelektronik Worldwide, 2017, 12(3): 18-23.

[12] Grad Ya A, Odinokov S B, Solomashenko A B, et al. Study of color reproduction features of AR device based on optical waveguides[C]//Optics, Photonics and Digital Technologies for Imaging Applications VI. Bellingham: Society of Photo-Optical Instrumentation Engineers, 2021: 11353.

[13] 倪一博, 闻顺, 沈子程, 等. 基于超构表面的多维光场感知[J]. 中国激光, 2021, 48(19): 233-260.

[14] Yu N, Genevet P, Kats M A, et al. Light propagation with phase discontinuities: Generalized laws of reflection and refraction[J]. Science, 2011, 334(6054): 333-337.

[15] 刘逸天, 陈琦凯, 唐志远, 等. 超表面透镜的像差分析和成像技术研究[J]. 中国光学, 2021, 14(4): 831-850.

[16] Li Z, Lin P, Huang Y W, et al. Meta-optics achieves RGB-achromatic focusing for virtual reality[J]. Science Advances, 2021, 7(5): eabe4458.

[17] Lee G Y, Hong J Y, Hwang S, et al. Metasurface eyepiece for augmented reality[J]. Nature Communications. 2018, 9(1): 4562.

[18] Zhang J L, Wang X R, Yang Y, et al. Flat dielectric metasurface lens array for three dimensional integral imaging[J]. Optics Communications, 2018, 414: 1-4.

[19] Deng H, Wang Q H. 3D display technology for augmented reality based on integral imaging-A review[J]. Science & Technology Review, 2018, 36(9): 18-24.

[20] Liu L, Ouyang W L, Wang X G, et al. Deep learning for generic object detection: A survey[J]. International Journal of Computer Vision, 2020, 128(2): 261-318.

[21] McBride T J, Vandayar N, Nixon K J. A comparison of skin detection algorithms for hand gesture recognition [C]//2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA). Piscataway NJ: IEEE, 2019: 211-216.

[22] 何胜皎. 视频序列中运动目标检测算法的研究[D]. 兰州: 兰州理工大学, 2018.

[23] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single rgb images[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 4903-4911.

[24] Gulati S, Bhogal R K. Comprehensive review of various hand detection approaches[C]//2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET). Piscataway, NJ: IEEE, 2018: 1-5.

[25] Zhao S Y, Yang W Y, Wang Y G. A new hand segmentation method based on fully convolutional network[C]//2018 Chinese Control and Decision Conference (CCDC). Piscataway, NJ: IEEE, 2018: 5966-5970.

[26] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent—A new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21271-21284.

[27] Liu Z P, Chai X J, Liu Z, et al. Continuous gesture recognition with hand-oriented spatiotemporal feature[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway, NJ: 2017: 3056-3064.

[28] Narayana P, Beveridge R, Draper B A. Gesture recogni-

tion: Focus on the hands[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5235-5244.

[29] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 779-788.

[30] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[31] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv Preprint, 2018: 1804.02767.

[32] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv Preprint, 2020: 2004.10934.

[33] Huu P N, The H L. Proposing recognition algorithms for hand gestures based on machine learning model[C]// 2019 19th International Symposium on Communications and Information Technologies (ISCIT). Piscataway, NJ: IEEE, 2019: 496-501.

[34] Zhan F. Hand gesture recognition with convolution neural networks[C]//2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). Piscataway NJ: IEEE, 2019: 295-298.

[35] Du T, Ren X M, Li H C. Gesture recognition method based on deep learning[C]//2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC). Piscataway, NJ: IEEE, 2018: 782-787.

[36] Hong J Y, Park S H, Baek J G. Segmented dynamic time warping based signal pattern classification[C]//2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). Piscataway, NJ: IEEE, 2019: 263-265.

[37] Tölgyessy M, Dekan M, Chovanec L′, et al. Evaluation of the azure Kinect and its comparison to Kinect V1 and Kinect V2[J]. Sensors, 2021, 21(2): 413.

[38] Mor B, Garhwal S, Kumar A. A systematic review of hidden markov models and their applications[J]. Archives of Computational Methods in Engineering, 2021, 28(3): 1429-1448.

[39] Haid M, Budaker B, Geiger M, et al. Inertial-based gesture recognition for artificial intelligent cockpit control using hidden markov models[C]//2019 IEEE International Conference on Consumer Electronics (ICCE). Piscataway, NJ: IEEE, 2019: 1-4.

[40] Tang J, Cheng H, Zhao Y, et al. Structured dynamic time warping for continuous hand trajectory gesture recognition[J]. Pattern Recognition, 2018, 80: 21-31.

[41] 张建荣. 基于Kinect手势识别的虚拟环境体感交互技术研究[D]. 重庆: 重庆邮电大学, 2016.

[42] Saha S, Lahiri R, Konar A, et al. HMM-based gesture recognition system using Kinect sensor for improvised human-computer interaction[C]//2017 International Joint Conference on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2017: 2776-2783.

[43] Khan A, Sohail A, Zahoora U, et al. A survey of the recent architectures of deep convolutional neural networks [J]. Artificial Intelligence Review, 2020, 53(8): 5455-5516.

[44] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: 2020: 203-213.

[45] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]//Proceedings of the IEEE Conference on Computer vision and Pattern Recognition. Piscataway, NJ: 2016: 1933-1941.

[46] Jing L L, Tian Y L. Self-supervised visual feature learning with deep neural networks: A survey[J]. IEEE Transactions On Pattern Analysis and Machine Intelligence, 2020, 43(11): 4037-4058.

[47] Han T, Xie W, Zisserman A. Self-supervised co-training for video representation learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 5679-5690.

[48] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 6202-6211.

[49] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of the 36th International Conference on Machine Learning. Brookline, MA: Microtome Publishing, 2019: 6105-6114.

[50] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 7083-7093.

[51] Yan S, Xiong Y, Lin D. Spatial temporal graph convolu-

tional networks for skeleton-based action recognition[C]. Thirty-second AAAI Conference on Artificial Intelligence. Washington: AAAI, 2018.

[52] Zhou H, Zhou W G, Zhou Y, et al. Spatial-temporal multi-cue network for continuous sign language recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2020, 34(7): 13009-13016.

[53] Zhu G M, Zhang L, Mei L, et al. Large-scale isolated gesture recognition using pyramidal 3D convolutional networks[C]//2016 23rd International Conference on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2016: 19-24.

[54] Li Y N, Miao Q G, Tian K, et al. Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28 (10): 2956-2964.

[55] Tran D, Ray J, Shou Z, et al. Convnet architecture search for spatiotemporal feature learning[J]. arXiv Preprint, 2017: 1708.05038.

[56] Miao Q G, Li Y N, Ouyang W L, et al. Multimodal gesture recognition based on the ResC3D network[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway, NJ: IEEE, 2017: 3047-3055.

[57] Li Y N, Miao Q G, Qi X D, et al. A spatiotemporal attention-based ResC3D model for large-scale gesture recognition[J]. Machine Vision and Applications, 2019, 30(5): 875-888.

[58] Gupta P, Kautz K. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway NJ: IEEE, 2016: 4207-4215.

[59] Zhu G M, Zhang L, Shen P Y, et al. Multimodal gesture recognition using 3D convolution and convolutional LSTM[J]. IEEE Access, 2017, 5: 4517-4524.

[60] Black M H, Chen N, Iyer K K, et al. Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography [J]. Neuroence and Biobehavioral Reviews, 2017, 80: 488-515.

[61] Moruzzi G, Magoun H W. Brain stem reticular formation and activation of the EEG[J]. Electroencephalography and Clinical Neurophysiology, 1949, 1(1/2/3/4): 455-473.

[62] Gibbs F A, Lennox W G, Gibbs E L. The electro-encephalogram in diagnosis and in localization of epileptic seizures[J]. Arch NeurPsych, 1936, 36(6): 1225-1235.

[63] Gabor A J, Seyal M. Automated interictal EEG spike detection using artificial neural networks[J]. Electroencephalography and Clinical Neurophysiology, 1992, 83(5): 271-280.

[64] Taran S, Bajaj V. Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method[J]. Computer Methods and Programs in Biomedicine, 2019, 173: 157-165.

[65] Chen L L, Zhang J, Zou J Z, et al. A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection[J]. Biomedical Signal Processing & Control, 2014, 10: 1-10.

[66] 张涛, 陈万忠, 李明阳. 基于频率切片小波变换和支持向量机的癫痫脑电信号自动检测[J]. 物理学报, 2016 (3): 038703.

[67] 邹凌, 王新光. 独立分量分析结合小波去噪算法提取诱发电位信号的仿真实验[J]. 中国组织工程研究, 2009, 13(43): 8503-8505.

[68] 李冬梅. 经验模式分解与代价敏感支持向量机在癫痫脑电信号分类中的应用[J]. 生物医学工程研究, 2017, 36(1): 33-37.

[69] 贺王鹏, 杨琳, 王芳, 等. 基于TQWT的癫痫脑电信号的识别[J]. 生物医学工程研究, 2017, 36(4): 346-350.

[70] Pachori R B, Bajaj V. Analysis of normal and epileptic seizure EEG signals using empirical mode decomposition [J]. Computer Methods and Programs in Biomedicine, 2011, 104(3): 373-381.

[71] 张发华, 舒琳, 邢晓芬. 头皮脑电采集技术研究[J]. 电子技术应用, 2017, 43(12): 3-8.

[72] 丁超. 便携式脑电采集系统设计[D]. 成都: 电子科技大学, 2013.

[73] 刘屏. 精神创伤后应激障碍及其防治研究进展[J]. 中国药物应用与监测, 2017, 14(1): 1-5.

[74] Pandey P, Seeja K R. Subject independent emotion recognition from EEG using VMD and deep learning[J]. Journal of King Saud University—Computer and Information Sciences, 2022, 34(5): 1730-1738.

[75] Camarda A, Salvia É, Vidal J, et al. Neural basis of functional fixedness during creative idea generation: An EEG study[J]. Neuropsychologia, 2018, 118(Part A): 4-12.

[76] Wankhade S B, Doye D D. IKKN predictor: An EEG sig-

nal based emotion recognition for HCI[J]. Wireless Personal Communications, 2019, 107(3): 12-15.
[77] Sutter E E. The brain response interface: Communication through visually-induced electrical brain responses [J]. Journal of Microcomputer Applications, 1992, 15(1): 31-45.
[78] Farwell L, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials[J]. Electro-encephalography and Clinical Neurophysiology, 1989, 70(6): 510-523.
[79] Pfurtscheller G, Silva F. Event-related EEG/MEG synchronization and desynchronization: Basic principles[J]. Clinical Neurophysiology, 1999, 110(11): 1842-1857.
[80] Wolpaw J R, McFarland D J, Neat G W, et al. An EEG-based brain-computer interface for cursor control[J]. Electroencephalography and Clinical Neurophysiology, 1991, 78(3): 252-259.

Metaverse terminal: Key hard technology development trends in virtual (augmented) reality

WANG Lijun, LI Zhengping, LI Ying, HOU Yaohui, WANG Jingliang, WANG Chuang, XU Zhiping, JIA Kehao, LIU Yuning
MA Weizhi

School of Information Science and Technology, North China University of Technology, Beijing 100144, China

Abstract With the development of the information technology era, metaverse has gradually become the third internet revolution, and the metaverse terminal is an important component of it. This article describes the development trends and key technologies of the hardware part of the metaverse terminal in terms of micro displays, optical systems, and perceptual interactions. In terms of the micro display technology, two technologies, silicon based OLED and MicroLED, are mainly discussed; in terms of optical systems, the main topics discussed are surface relief grating waveguide technology, volume holographic waveguide technology, and metasurface waveguide technology; in terms of perceptual interaction, the interaction technology of gesture recognition and EEG recognition is mainly discussed.

Keywords metaverse; virtual reality (VR); augmented reality (AR); microdisplays; perceptual interaction; artificial intelligence

（责任编辑　王志敏）