虚拟现实的技术瓶颈

曹煊[1,2]

1. 中国科学院自动化研究所,北京 100190
2. 南加州大学ICT实验室,美国洛杉矶 12015

摘要 在技术变革和资本力量的双重推动下,虚拟现实(virtual reality)技术在近几年发展迅速,初步达到了可商业化的程度。虚拟现实和3D电影院都是通过双目视差实现三维成像,但虚拟现实提供了3D电影院所不具备的移动视差并提供了强烈的沉浸感。现阶段虚拟现实技术仍面临着一系列技术难题,其中眩晕和人眼疲劳尤其明显,是虚拟现实的技术禁地。本文从介绍三维视觉感知开始,分析了虚拟现实造成眩晕和人眼疲劳的根本原因。给出了解决这一技术瓶颈的答案——动态光场,并从光场采集和显示两方面分析了多种光场技术的优缺点。列举了增强现实(augmented reality)技术的3种实现形式,并从人与人交互和通信的角度对比了虚拟现实与增强现实在未来的发展趋势。

关键词 虚拟现实;增强现实;动态光场;计算摄像;数字全息显示

1 虚拟现实一直存在

近几年,虚拟现实(virtual reality,VR)技术发展迅猛,商业化、市场化和产品化的趋势日益明显。然而,早在50多年前,科学家们就已经提出了虚拟现实的技术构想。美国计算机图形学之父Ivan Sutherland 在1968年开发了第一个图形可视化的“虚拟现实”设备,但在当时还不叫“虚拟现实”,而是被称为“头戴显示”或“头盔显示”(head-mounted display,HMD)。就技术层面而言,现阶段的虚拟现实眼镜或者虚拟现实头盔仍可划分为HMD的范畴。

2013年谷歌眼镜(Google Glass)面市,“虚拟现实”这个术语开始进入公众视野。但当时的谷歌眼镜没有双目立体视觉,所以称为Google Glass,而不是Google Glasses。尽管谷歌眼镜的整体显示效果低于同一时期的手机和计算机,但其新颖的成像方式引起了人们的极大关注。这背后揭示了人们对于已经沿用了20多年的传统平面显示方式的审美疲劳和对新颖显示方式的强烈期待。

总体来说,现阶段虚拟现实有三大显著特点(简称为3I):沉浸感(Immersion)、交互性(Interaction)和构想性(Imagination)。视觉是人类最敏感,捕获信息量最大的“传感器”。VR眼镜隔绝了人眼接收外部视觉信息的通道,取而代之的是虚拟的视觉内容。当人眼受到来自VR眼镜的视觉刺激时,大脑会自动“绘制”出虚拟的环境,从而使人沉浸在一个全新的环境中。相比于传统的显示方式,交互性并不是VR所特有的。电视可以借助遥控器交互,计算机可以借助鼠标键盘输入。目前虚拟现实还没有统一的输入设备,交互方式可以根据虚拟场景来设置,更具灵活性和多样性。例如在士兵培训中,VR交互方式可以是一把枪;在模拟外科手术中,交互方式可以是手术刀。人们借助VR可以以第一人称视角探索未知的环境,包括一些人类难以到达的环境,例如深海、外太空;甚至包括一些人类无法到达的或抽象的环境,例如细胞、黑洞、数学模型。VR技术给了人们一个可以徜徉在任何环境中的机会。在这样一个从未到达的环境中,人类的视野和想象力得到了极大的延展。

既然虚拟现实早就存在,但为什么直到现在才爆发呢?一方面是因为虚拟现实作为一种全新的显示方式,正好满足了人们对于信息可视化变革的期待。另一方面也是因为技术变革和资本力量的共同驱动。

2 VR背后的支撑

在此之前,大规模普及虚拟现实还只是一个美丽的梦,因为受到计算性能、工业集成化、可视化技术发展的限制。而近10年来,相关的技术得到了迅猛的发展,为VR的商业化和产品化奠定了技术基础。除此之外,有一股不可忽视的力量在推动VR加速发展,那就是大资本。

2.1 VR背后的技术变革

显示技术的发展可以划分为4个阶段:平面2D→曲面2.5D→头戴显示3D→裸眼全息。人类生存的世界是三维的,但自从相机和显示器诞生以来,一直以二维平面的方式记录和显示这个三维世界,这是一种降维后的表现方式。从早期的阴极射线管显示器(CRT)到轻薄的液晶显示器(LCD),从

收稿日期:2016-06-15;修回日期:2016-06-30

作者简介:曹煊,博士研究生,研究方向为裸眼三维显示、光场的采集和显示、计算摄像、虚拟现实及增强现实,电子信箱:caoxuan21@126.com

引用格式:曹煊. 虚拟现实的技术瓶颈[J]. 科技导报, 2016, 34(15): 94-103; doi: 10.3981/j.issn.1000-7857.2016.15.013

黑白显示到彩色显示,每一次技术变革都没有突破显示维度的限制。全世界的科学家们都在努力尝试打破这一困境,试图还原一个真实的3D世界。在虚拟现实技术出现在公众视野之前,有另外两种突破二维显示的技术出现在了消费市场,包括曲面2.5D显示和裸眼3D显示,但这两种技术都未能获得消费者的"芳心"。曲面2.5D显示技术并没有带来信息可视化在维度上的突破,人们并不能从该显示器中感知到第三维度的信息(视觉深度感)。裸眼3D显示技术为观看者带来了视觉深度感,但目前的裸眼3D显示技术还存在很多的技术难点有待突破,包括分辨率损失严重、观看视角狭窄、相邻视点跳跃等。在可预见的未来,裸眼3D技术还无法达到令消费者满意的效果。因此,上述两种超二维显示技术都未能调和技术可行性与市场期待之间的矛盾。在这样的局面下,虚拟现实应运而生,它是技术可行性和市场期待的折中产物。

2.2 VR背后的资本力量

除了相关技术的变革和发展,资本力量的推动也是VR蓬勃发展的另一重要因素。如果说2013年谷歌眼镜的推出是行业大鳄窥视头戴显示巨大宝藏的一隅,那么2014年Facebook收购Oculus就是巨大资本撬开虚拟现实潘多拉魔盒的开始(注:Oculus是一家专注于虚拟现实技术的公司)。随着资本的进入,更多的科研力量、工程技术以及3D内容开发都纷纷进入了该领域。2016年被称为虚拟现实元年,HTC、Facebook、Sony等国际巨头,以及国内的部分虚拟现实公司都将自己的VR产品正式推向了市场。在这样的国际格局下,国内的部分资金也开始疯狂投向虚拟现实领域。

3 为什么能感知到三维

人们生活的世界是一个四维空间,包括水平维度、垂直维度、纵深维度和时间维度。例如在图书馆寻找一本书需要知道书籍处于第几排、第几列的书架,以及处于书架的第几层。并且还需要知道这本书是否已经借出,什么时候会出现在该书架。通过视觉观察物理世界时具有即时性,一般假设光线从环境中发出到人眼接收的时间为零,因此不用考虑时间维度,用前三个维度描述所观察的世界。例如伸手拿杯子时,视觉系统会帮助判断杯子处于手的左边还是右边,上边还是下边,前面还是后面。在一个平面上可以很容易地感知到水平维度和垂直维度,但如何感知到第三维度——视觉深度呢?

众所周知,双目视差是提供视觉深度的重要途径,但视觉深度不仅仅由双目差体现,单眼也能感知到深度。深度信息(depth cues)有很多种[1],主要包括以下信息。

1) 双目视差(binocular parallax),也称为左右视差或双目汇聚。所观察的物体越近,视差越大(图1),双眼汇聚角度越大(图2);所观察的物体越远,视差越小,双眼汇聚角度越小。必须依靠双目协同工作才能感知到双目视差。

2) 移动视差(motion parallax),当观察视点改变后,远近不同的物体在人眼中产生的位移会不同,如图3所示。经过相同的视点改变,远处的物体在人眼中产生的位移更小,近处的物体在人眼中产生的位移更大。双目和单目都可以感知到移动视差。

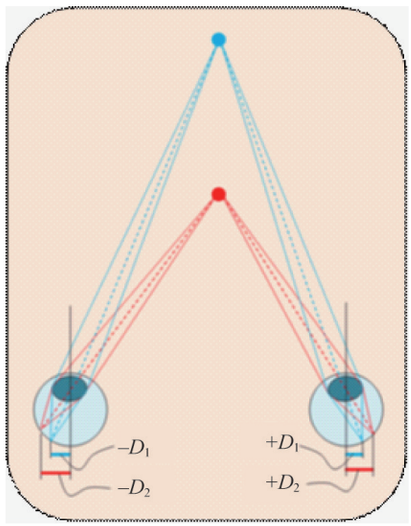

图1 双目视差

Fig. 1 The binocular parallax

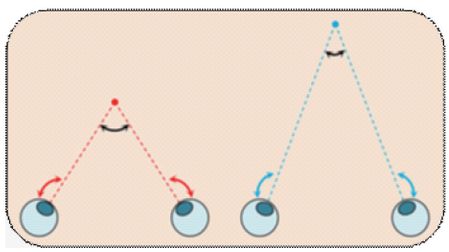

图2 双目汇聚

Fig. 2 The binocular convergence

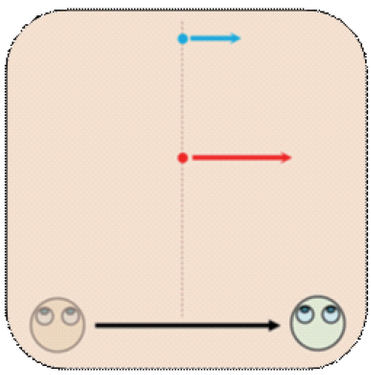

图3 移动视差

Fig. 3 The motion parallax

3）聚焦模糊(focus-blur),人眼的工作原理可以简化为一个照相机。当改变相机镜头的焦距时,相机可以聚焦在远近不同的平面上,从而使聚焦平面上的物体清晰成像,非聚焦平面的物体成像模糊。人眼的睫状肌就扮演着“相机镜头”的角色。如图4所示,当睫状肌紧绷时,人眼聚焦在近处平面;当睫状肌舒张时,人眼聚焦在远处平面。根据睫状肌的屈张程度,视觉系统可以判断出物体的相对远近。单目即可明显感知到聚焦模糊。

除了上述3种主要的深度信息,大脑会根据一些视觉经验来判断物体远近,例如遮挡关系、近大远小关系;同时也会根据一些先验知识作为辅助判断,例如看到一个杯子,先验知识会告诉大脑杯子不会太远;若看到一座高山,先验知识会告诉大脑高山在很远的地方。

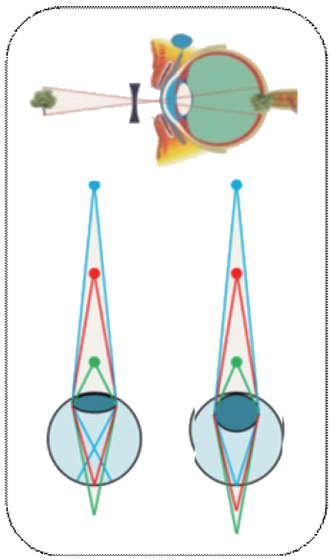

图4　聚焦模糊

Fig. 4　The focus-blur

4　VR的基本原理

虚拟现实的三维成像原理并不复杂,其基本原理和3D电影院一致,如图5所示,都是给左右眼分别呈现不同的图像,从而产生双目视差。当大脑在合成左右眼的图像时,会根据视差大小判断出物体的远近[1]。虚拟现实眼镜不仅提供了双目视差,还提供了3D电影院所不具备的移动视差信息。当坐在3D电影院的第一排最左边和最右边的位置时,所看到的3D内容是一样的。但正确的3D成像方式应该是:坐在最左排的观看者看见物体的左侧面,坐在最右排的观看者看见物体的右侧面。例如观看桌面上的茶杯时,左右移动头部会看见茶杯的不同侧面。如图6所示,虚拟现实眼镜同时提供了双目视差和移动视差,不仅左右眼图像不同,而且当旋转或平移头部时看见的3D内容也不同。

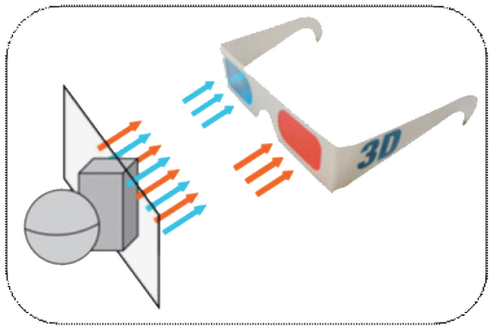

图5　3D电影院成像原理

Fig. 5　Principle of 3D cinema

图6　虚拟现实头戴显示设备Oculus Rift

Fig. 6　VR glasses Oculus Rift

(图片来源:Oculus官方网站)

当前VR产品形态主要分为3种:基于手机的VR、VR一体机、基于PC机的VR,主要特点如表1所示。由于技术和成本的限制,当前的VR产品都在价格、性能、舒适度三者之间平衡,上述3种形态的VR产品只是在不同的方面有所侧重。目前消费市场中尚未出现低价格、高性能的轻薄VR眼镜。同时从表1中也可以看出,从低廉的到昂贵的VR产品都会引起眩晕和人眼疲劳。高性能的VR产品在眩晕的耐受时间上稍微有所延长,但仍然无法达到像智能手机一样长时间使用。

表1　当前VR产品形态

Table 1　Current VR products

VR产品形态	价格	运算性能	续航时间	佩戴	VR体验	眩晕
基于手机的VR	低廉百元级	中等嵌入式等级	中等取决于手机	较重	一般	是
VR一体机	昂贵千元级	中等嵌入式等级	较短数小时	沉重	一般	是
基于PC机的VR	总价最贵需高配置PC	较高取决于PC	无限时长	较轻	较好	是

虚拟现实根据使用场景大致可以分为座椅式、站立式,场地式。顾名思义,座椅式VR限制用户位在座椅上,只能检测到视点的姿态旋转变换(Pitch,Yaw,Roll),而忽略视点平移变化。如图7所示,Pitch围绕x轴旋转,也叫做俯仰角,Yaw是围绕y轴旋转,也叫偏航角,Roll是围绕z轴旋转,也称翻滚角。而站立式VR和场地式VR都能同时检测到视点的姿态旋转变化和平移变化。站立式VR允许用户在独立的房间内(一般为10 m×10 m以内)自由走动,活动范围较狭窄,不适用于模拟大范围的场景。场地式VR理论上允许用户可以在无限范围内自由走动,是真正意义上的虚拟世界。但鉴于场地有限,传感器的工作范围有限。实际中场地式VR需要万向跑步机的支撑,将跑步机履带的平移数据转化为人体的移动数据。表2中所列举的交互方式是对应场景下的主要交互方式而非唯一交互方式。目前虚拟现实还没有标准的输入设备。在传统手柄的基础上,出现了一些新颖的VR输入方式。头控是指通过头部的运动改变指针位置,通过悬停表示确认。线控是指通过现有的连接线(例如耳机线)实现简单的按键操作。触摸板一般位于VR头盔的侧面,与笔记本电脑的触摸板实现相同的功能。根据VR场景,交互方式也可以是仿手型手柄,例如枪械、手术刀等。

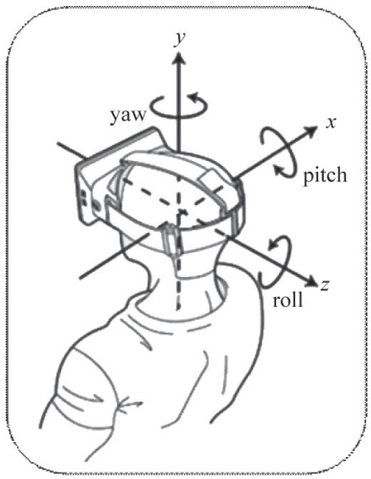

图7　头部姿态变化的3个自由度

Fig. 7　Three degrees of freedom of the head movement

(图片来源:Oculus Rift SDK文档插图)

表2　当前VR使用场景

Table 2　Current VR application scenarios

VR应用场景	视点自由度	活动范围	交互	定位
座椅式	三轴旋转	半径1 m	手柄为主头控,线控,触摸板	头部角度定位
站立式	三轴旋转和平移	独立房间	仿手型手柄为主例如枪械、手术刀	头部角度和平移定位 手柄角度和平移定位
场地式	三轴旋转和平移	理论上可无限扩展	根据应用而定需要万向跑步机	头部角度和平移定位 手柄角度和平移定位

5　VR的技术瓶颈

虚拟现实技术经过近几年的快速发展,各方面性能逐步完善,但仍然面临着一些关键技术有待改进和突破。主要可以概括为下列3个方面。

1)大范围多目标精确实时定位。

目前在已经面向市场的VR产品中,当属HTC Vive Pre的定位精度最高,时延最低。HTC Vive Pre的定位主要依靠Light House来完成。Light House包括红外发射装置和红外接收装置。红外发射装置沿着水平和垂直两个方向高速扫描特定空间,在头盔和手柄上均布有不少于3个红外接收器,且头盔(手柄)上所有的红外接收器之间的相对位置保持不变。当红外激光扫过头盔或手柄上的红外接收器时,接收器会立即响应。根据多个红外接收器之间的响应时间差,不仅可以计算出头盔(手柄)的空间位置信息还能得出姿态角度信息。目前HTC Vive Pre只能工作于一个独立的空旷房间中。障碍物会阻挡红外光的传播。而大范围、复杂场景中的定位技术仍需突破。多目标定位对于多人同时参与的应用场景至关重要。当前的虚拟现实系统主要为个人提供沉浸式体验,例如单个士兵作战训练。当多个士兵同时参与时,彼此希望看见队友,从而到达一种更真实的群体作战训练,这不仅需要对多个目标进行定位,还需要实现多个目标的数据共享。

2)感知的延伸。

视觉是人体最重要、最复杂、信息量最大的传感器。人类大部分行为的执行都需要依赖视觉,例如日常的避障、捉取、识图等,但视觉并不是人类唯一的感知通道。虚拟现实所创造的模拟环境不应仅仅局限于视觉刺激,还应包括其他的感知,例如触觉、嗅觉等。

3)减轻眩晕和人眼疲劳。

目前所有在售的VR产品都存在导致佩戴者眩晕和人眼疲劳的问题。其耐受时间与VR画面内容有关,且因人而异,一般耐受时间为5~20 min;对于画面过度平缓的VR内容,部分人群可以耐受数小时。

上述的技术瓶颈中,大范围多目标精确实时定位已经取得了一定的突破,在成本允许的情况下,通过大面积的部署传感器是可以解决这一问题的。感知的延伸还存在较大的

技术难度,尤其是触觉;但当前的VR应用对感知的延伸并没有迫切的需求。相比之下,眩晕和人眼疲劳却是一个到目前为止还没有解决但又迫切需要解决的问题,是现阶段虚拟现实的技术禁地。

5.1 为什么会眩晕?

虚拟现实比3D电影提供了更丰富的三维感知信息,更逼近于人眼观看三维物理世界的方式。但为什么VR眼镜在佩戴一段时间后会导致眩晕和人眼疲劳呢?其原因是多样的,主要包括如下3方面。

1)身已动而画面未动。

如果无法获取VR眼镜的姿态和平移信息,则无法感知到移动视差。身体移动后,观看视点的位置和观看角度也随之改变,但人眼看见的3D画面并没有相应的改变。这会导致大脑在处理视觉信息和肢体运动信息时产生冲突,从而在一定程度上导致眩晕不适。

2)画面已动而身未动。

目前虚拟现实的应用还局限在一个非常有限的物理空间内。当画面快速变化时,身体的运动也应该与之匹配,但受到运动范围的限制,身体并没有产生对应幅度的运动,从而在大脑中产生了肢体运动信息和视觉信息的冲突。例如,通过虚拟现实体验过山车时,观看视点和角度在快速地变化,但身体却保持不变。当VR画面变化(过度)越快时,大脑产生的冲突越明显。

上述两种眩晕都是由视觉信息与肢体运动信息之间的冲突造成的,统称为晕动症。产生晕动症的技术原因是多方面的。

(1)空间位置定位和姿态角度定位的精度和速度。惯性测量装置(inertial measurement unit,IMU)是一种微机电(MEMS)模块,也是当前VR眼镜测量角度姿态的主要技术手段。但IMU只能测量姿态角度,不能测量空间位移。多个IMU组合可以实现空间位移测量,但积累误差大且难以消除,暂不适用于VR眼镜。另一种定位技术是基于传统摄像头的SLAM(simultaneous localization and mapping)算法[2],可以同时实现空间位置定位和姿态角度定位且适用于复杂场景,但目前SLAM算法在精度、速度和稳定性上都有待提高。基于双目相机或深度相机的SLAM是一个有价值的潜在研究方向。目前最实用的定位技术是HTC Vive Pre中应用的红外激光定位技术,硬件成本低且同时具备高精度低时延的空间位置定位和姿态角度定位,但其应用局限于小范围的空旷场景中。

(2)显示器件的刷新频率。目前头戴显示(HMD)的像源主要包括微投影仪和显示屏两种。其中微投影仪主要应用在增强现实(AR,Argumented Reality)中,例如Google Glass,Hololens,Meta,Lumus,Magic Leap等。虚拟现实主要采用小尺寸显示屏(6寸以下)作为像源,其中显示屏又分为液晶显示屏(LCD,Liquid Crystal Display)和有机自发光显示屏(OLED,organic light-emitting diode)。目前LCD和OLED屏幕的刷新率普遍能达到60 Hz以上,部分型号甚至能达到90 Hz以上。OLED采用自发光成像,因此余晖比LCD更小,上一帧图像的残影更小。

(3)图像渲染时延。虚拟现实所创建的模拟环境是经计算机图形图像学渲染生成得到。渲染的速度直接由计算机性能决定,尤其依赖于计算机中的显卡(graphic processing unit,GPU)性能。目前高性能的GPU渲染一个复杂场景已能达到全高清(Full HD)90 fps以上。

VR眼镜的图像刷新速度取决于上述3个技术指标的最低值。也即,上述3个环节中,任何1个环节速度慢都会导致图像刷新率降低,从而出现晕动症。在前几年,VR设备厂商将VR眼镜的眩晕归因于"图像刷新太慢"。目前最新的VR眼镜在空间位置定位和姿态角度定位的速度、显示器件的刷新频率、图像渲染速率3个指标均能达到90 Hz,远高于人眼时间暂留的刷新阈值(24 Hz)。为什么还是会眩晕呢?有人怀疑是活动范围有限导致身体移动的幅度与画面变化幅度不一致。万向跑步机无限延伸了活动范围,但眩晕的问题依然存在。由此可见,上述两个方面是造成了眩晕的表象原因,并不是根本原因。

3)聚焦与视差冲突。

对于双目视差、移动视差、聚焦模糊3种主要深度信息,当前的头戴显示设备只提供了前两种。聚焦丢失(聚焦错乱)是产生眩晕的"罪魁祸首"。"聚焦模糊"真的就这么重要吗?众所周知,双眼能感知物体远近,但其实单眼也可以。当伸出手指,只用一只眼注视手指时,前方的景物模糊了;而当注视前方景物时,手指变的模糊,这是由眼睛的睫状肌屈张调节来实现的。眼镜聚焦在近处时,睫状肌收缩,近处的物体清晰而远处的场景模糊;眼镜聚焦在远处时,睫状肌舒张,远处的场景清晰而近处的物体模糊。通过睫状肌的屈张程度能粗略感知到物体的远近,因此单眼也能感知到立体三维信息。如图8所示,现阶段的虚拟现实头显设备只提供单一景深的图片,且图片的景深固定。这导致人眼始终聚焦在固定距离的平面上。当通过"聚焦模糊"感知到的深度信息与通过"双目视差"感知到的深度信息不一致时,就会在大脑

图8 现阶段的虚拟现实头显设备只提供单一景深画面

Fig. 8 With the current VR glasses, only images of one depth of focus can be seen

(图片来源:http://www.yule.com.cn/html/201601/9744.html)

中产生严重的冲突,称为“聚焦与视差冲突”(accommodation-convergence conflict,ACC)[3~6]。而且当大脑检测到ACC时,会强迫睫状肌调节到新的屈张水平使之与双目视差所提供的深度信息相匹配。当睫状肌被强迫调节后,因为聚焦错乱,图像会变的模糊;此时大脑会重新命令睫状肌调节到之前的屈张水平。如此周而复始,大脑就“烧”了。

回到之前3D电影眩晕的问题,当观看者坐在第一排中间位置时,双眼到大荧幕距离为10 m且保持不变。当3D内容为远处的高山时,双目视差较小,会引导人眼注视于前方几百米处。而人眼接收的光线都来自10 m处的大荧幕,左眼和右眼会自主地聚焦在10 m处的平面上以便能清晰地看见图像。此时双目的汇聚和睫状肌的屈张水平不一致,从而导致了人眼不适。同理,当3D内容为眼前1 m处的一条蛇时,人眼仍然聚焦在10 m处的平面,从而产生类似的聚焦与视差冲突。

聚焦与视差之间的冲突比视觉信息与肢体运动信息之间的冲突更严重。举个例子,反恐精英(Counter-Strike,CS)是一款风靡世界的射击类游戏,玩家以第一人称视点在虚拟环境中奔跑,跳跃和射击。当画面变化时,玩家仍然静坐在计算机前,并没有实际的跑动和跳跃。此时玩家并没有产生眩晕的感觉,甚至能长时间沉浸其中。其原因在于玩家经过一段时间的训练以后,在大脑中建立了肢体运动与鼠标键盘操作之间的映射关系,比如前后左右跑动与键盘W、S、A、D按键对应,跳跃与空格按键对应。因此,通过运动关系的映射,视觉信息与肢体运动信息之间的冲突(晕动症)得以大大减轻,但睫状肌的屈张是一种自发行为。睫状肌会自主地屈张到正确的水平,以保证人眼聚焦在所关注物体的表面,并且人眼总是趋向于得到最清晰的视觉成像,这也会促使睫状肌处于与之匹配的屈张水平。因此强迫睫状肌处于非正确的屈张水平或被错误地引导到不匹配的屈张水平都会导致上述的冲突,从而导致眩晕和人眼疲劳。通过训练来建立类似于“反恐精英”中的大脑映射是无法解决此类冲突的,只能通过头戴显示设备产生不同深度的图片引导人眼自然地聚焦在远近不同的平面上才能从根本上解决这一冲突,从而解决眩晕和人眼疲劳。

VR眼镜的严重眩晕问题引发了对另一个问题的思考,为什么3D电影在数小时后才出现眩晕或人眼疲劳,而VR眼镜的耐受时间一般只有5~20 min? 一方面是因为3D电影已经普及多年,能适应3D电影的人群已经变得更加适应,不能适应3D电影的人群已经不再去3D电影院,所以造成所有人都能耐受3D电影数小时的假象。另一方面,3D电影是第三人称视角观看,而虚拟现实使观看者处于第一人称视角,晕动症更加明显。再一方面,3D电影的荧幕距离人眼较远(一般十米到几十米不等),虽然聚焦错乱的问题依然存在,但睫状肌始终处于较舒张的状态。而VR眼镜的屏幕经准直透镜放大以后,一般等效在较近处(一般2~5 m),睫状肌始终保持紧绷的状态,人眼更易疲劳。上述3个原因导致了虚拟现实的耐受时间相比于3D电影缩短了很多。

眩晕是目前虚拟现实最大的技术瓶颈,大大限制了虚拟现实产业的长足发展,并且会对人眼造成伤害。在VR眼镜佩戴的全过程中都会强迫人眼处于错误的聚焦平面,睫状肌得不到连续自然的舒张和收缩。长此以往,睫状肌弹性下降,失去了自主调节的能力,从而导致近视。尤其对于12岁以下的儿童,人眼器官正处于生长发育阶段,VR眼镜会大大增加患近视的可能性。即使是成年人,长期佩戴也会导致视力下降。因此虚拟现实应用于幼教领域需严格控制其佩戴时间。幼儿应尽可能减少甚至不佩戴VR眼镜,直到突破这一技术瓶颈。

5.2 光场显示技术

在讨论如何解决虚拟现实的眩晕问题之前,先思考人眼是如何观看三维物理世界的?

环境表面的每一个点都会在半球范围内发出光线(自发光或反射光)。空间中的点可以通过三维坐标(x,y,z)来唯一表示;每个点在半球范围内发出的光线通过水平夹角ϕ和垂直夹角φ来描述;光线的颜色通过波长λ表示(光线还包括亮度信息,这里用λ统一表示);环境光线随着时间是变化的,不同时刻t下的光线也不一样。因此,环境光线可以通过7个维度的变量来描述[7],称为全光函数$P(x,y,z,\phi,\varphi,\lambda,t)$(图9)。假设环境光线在一定时间内稳定不变,则每条光线的波长可以用5D函数表示为$\lambda=F(x,y,z,\phi,\varphi)$。

如果显示器能产生上述5D函数中所有的光线,则观看

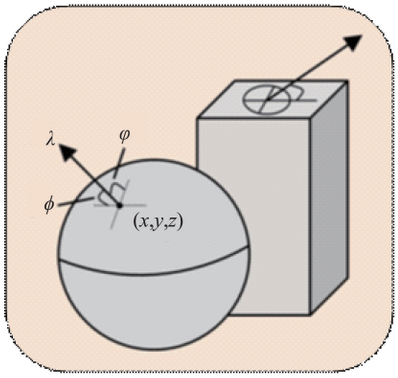

图9 全光函数模型

Fig. 9 A model of plenoptic function

者通过该显示器能在视觉上感知到与真实世界中一样的三维环境。但遗憾的是,目前全世界都没有这样的显示器。当前的电视、计算机、手机等平面显示屏只实现了上述5D函数中的2个维度,也即$\lambda=F(x,y)$。近几年出现的曲面显示屏增加了维度z上的像素点,但在维度z上并不完备。因此,曲面显示屏不是3D显示器,只能算作2.5D显示器。科学家们曾尝试了多种方法从传统的2个维度显示提升到更高维度显示,但目前仍停留在实验室阶段,尚无可商业化的产品。例如:1)体三维显示[8](Volumetric 3D Display)在空间中不同位

置发出光线,实现了 $F(x,y,z)$ 3个维度的显示,但依赖于机械运动,且无法呈现正确的遮挡关系;2) 基于微透镜阵列的集成成像[9](integral imaging)需要将一层特殊的光学膜贴在平面显示屏上,实现了 $F(x,y,\phi,\varphi)$ 4个维度的显示,但图像分辨率大大降低,且在 (ϕ,φ) 维度上采样率越高,图像的分辨率损失越严重;3) 投影仪阵列[10](Projector Array)从不同的方向发出不同的光线,实现了 $F(x,y,\phi,\varphi)$ 4个维度的显示且分辨率不损失,但硬件成本高昂且体积大。

如果能将传统的2D平面显示提升到5D显示,人眼将不借助任何头戴设备而获得类似全息显示的效果。但根据显示领域目前的技术发展,在未来较长一段时间内难以实现轻便低廉的5D全光显示器。

如图10所示,上述的5D全光函数是从"环境表面发出了什么光线"这一角度来建立数学模型。但从另一个角度来建模将会简化问题——"观看环境时,人眼接收了什么光线?"。如果头戴显示器能重现出人眼应该接收的全部光线,人眼将从头戴显示器中看到真实的三维场景。

5D全光函数描述了环境表面发出的所有光线,但并不是

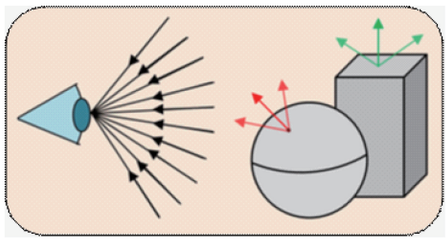

图10 环境表面发出的光线和人眼接收到的光线

Fig. 10 The light reflected by objects and that perceived by eyes

所有的光线都进入了人眼,只有部分光线最终被人眼接收。因此进入人眼的光线是5D全光函数的一个子集。且随着人眼位置和注视方向的不同,人眼接收到不同子集的光线。

将人眼的瞳孔分为 $N_x \times N_y$ 个子区,用 (x,y) 表示横向第 x 个,纵向第 y 个瞳孔子区,图11中左图展示了一个4×4瞳孔分区的视觉成像模型。如果瞳孔的分区 $N_x=1$, $N_y=1$;也即整个瞳孔作为一个区,这与传统的小孔成像模型是等效的。每个子区都会接收到很多从不同角度入射的光线,入射角度用 (α,β) 表示。因此,进入人眼的光线可以通过一个4D函数描述,可以称之为全视函数 $\lambda = F(x,y,\alpha,\beta)$。光线进入人眼的位置 (x,y) 和进入的角度 (α,β) 共同决定了光线会落在视网膜上的什么位置。如果不考虑与眼睛注视方向垂直的光线,5D全光函数可以降维到4D光线集合,一般用两个平面 (u,v) 和 (s,t) 表示,称为"光场"[11]。本文中采用一个平面 (x,y) 和一对角度 (α,β) 表示人眼接收光线的集合,是一种更适合于头戴显示的光场定义。

头戴显示设备如何投射出4D光场呢?假设光线在传播

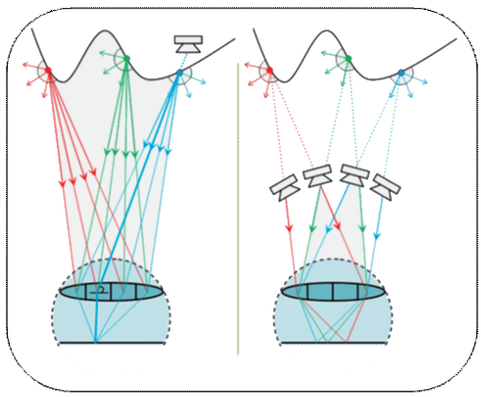

图11 全视函数模型

Fig. 11 A model of all vision function

过程中被看做一条射线,且沿着射线的方向上亮度和颜色不改变。例如图11左图中蓝点发出的第二条光线(蓝色粗线)与其射线方向上投影仪发出的光线是等效的,这样的假设对于日常环境中的光线传播完全合理。

基于上述合理假设,采用投影仪阵列可以模拟重现出4D光场,如图11中右图所示。当投影仪足够多、足够密集时,就可以在一定视野范围内无限逼近地投射出人眼应该接收到的全部光线。但投影仪体积较大,无法密集排列,且硬件成本高。值得一提的是,美国Magic Leap公司在2015年展示了一种基于光纤微型投影仪阵列的动态光场成像技术,大大减小了投影仪阵列的体积,提高了投影仪排列密度,但硬件成本仍然高昂。

投影阵列通过增加显示器件提高成像维度,这是一种最直接的将传统2D显示提升到4D光场显示的方法。但是通过不断增加硬件设备增加像源的自由度并不是一种高效的解决方案。首先硬件成本会急剧增加,例如实现图11右图中4×4投影阵列的光场,需要16倍的硬件成本;且数据的存储和传输也会增加到16倍。

光场显示为什么能解决头戴显示的眩晕问题呢?如上所述,光场显示提供了真实环境中发出的并由人眼接收的全部光线。人眼在观看真实环境时不眩晕,那么通过光场头显设备也就不会眩晕。如图4中,远近不同的点进入人眼的角度不同,这在4D光场 $\lambda = F(x,y,\alpha,\beta)$ 中通过角度参数 (α,β) 体现。因此,通过光场显示,人眼能自然的聚焦在远近不同的发光点上。从而睫状肌的屈张水平始终与双目视差保持一致,避免在大脑中产生ACC冲突。如图12所示,当同时呈现远近不同的图像层时,人眼能够自主地选择聚焦平面。真实环境中,图像层数达到无穷多层,由近及远连续分布。这意味着需要无穷多台投影仪才能重现连续分布的图像层,这显然是不切实际的。因此,在实际的光场显示中采用离散的图像层去近似逼近连续的图像层。当图像层数达到8层及以上时,人眼就能获得近似的聚焦感知。当然,图像层数越多,聚

焦越连续,视觉效果越自然,眩晕改善越显著。当前所有在售的头戴显示设备都只提供了1层图像,还远远不能达到近似连续聚焦的成像效果。

除了投影阵列,还有多种技术可以实现光场显示。例如,时分复用的投影技术采用一台高速投影仪从空间中不同位置投射图像,通过复用一台高速投影仪去"顶替"投影仪阵列[12]。但目前实现微型化的高精度机械控制比较困难,因此该技术不适用于头戴显示。断层成像[13,14]技术实现了数字化的空间光调制,只需要2~3倍的硬件成本就能实现5×5的光场成像,但计算量大、算法复杂度高,当前的个人计算机还无法实现在线高分辨的光场计算。该技术适用于离线应用(如光场电影)或者可在云端计算完成的应用(如光场虚拟现实直播)。

综上所述,光场是最接近人眼观看自然环境的成像方式,弥补了当前头戴显示都不具备的"聚焦模糊",将人眼睫状肌从固定的屈张水平中解放出来,消除了眩晕,减轻了人眼疲劳。实现光场成像已有多种技术手段,但都有各自的缺陷。受成本、计算量、设备体积的限制,当前的光场成像技术还只能在部分行业应用。

目前在售的VR眼镜普遍都比较厚重,轻薄化是虚拟现实设备未来的必然趋势。可以通过优化光学设计,减小透镜的焦距来缩短光程,从而减小VR眼镜的厚度,但短焦距的透镜会带来色差和畸变等其他光学问题,且透镜重量会随着焦距的缩短而增加。光场成像不仅解决了眩晕问题,还能使头显设备变得更轻更薄。基于上述光线在射线传播方向上具有不变性的假设,投影阵列可以移动到更靠近眼睛的位置,在不改变透镜焦距的前提下可以缩短光程,只需要根据投影阵列与透镜的相对位置对光线进行反向追迹渲染即可获得等效的光场成像。

最近出现了一些基于眼球追踪的光场显示技术,其根据人眼的注视方向,选择性的模糊掉人眼并不关注的像素块,从而造成一种人眼可以主动选择聚焦的假象。这一类技术可以归为伪光场成像。究其本质,伪光场成像技术仍然只提供了 $\lambda = F(x,y)$ 两个维度上的光线。换言之,伪光场成像技术只提供了1层图像,人眼仍然无法主动选择性聚焦,眩晕的问题依然没有得到解决。

5.3 计算摄像

光场成像技术显示了4个维度的光线,但如何采集4D光线呢?在计算机中可以对三维模型直接渲染得到4D光场,但是如何拍摄真实场景中的4D光场呢?可以明确的是,传统的摄像技术是无法采集4D光场的。摄像技术最早可以追溯到小孔成像,现今使用的相机仍然沿用着小孔成像模型。如图12所示,光场成像技术在不同深度上呈现多幅图片。而传统的相机只在一个聚焦平面上采集图像。传统相机拍摄的平面2D图片只是4D光场的一个子集。因此大量的光线信息在拍摄过程中丢失了。要显示光场,首先要解决如何采集光场的问题,否则"巧妇难为无米之炊"。

光场采集依赖于一门称为计算摄像(computational photography)的学科。最早的计算摄像是基于大量的相机从不同的角度分别拍摄采集光场,也称之为相机阵列[15,16]。当然也可以采用单个相机移动拍摄,但只能采集静态场景的光场。相机阵列是早期形态的光场相机,占地面积大,操作复杂,成本昂贵。目前市面上已经出现了消费级的光场相机(如Lytro[17])可以在单次拍摄中采集光场。Lytro光场相机采用微透镜阵列(microlens array)采集不同角度入射的光线。相比于相机阵列,Lytro光场相机体积大大减小,硬件成本降低,但分辨率也大大降低。基于上述两种光场相机的优缺点,科学家们提出了一种基于压缩感知的光场相机[18,19]。该光场相机通过"学习"已采集的光场,训练得到光场字典。利用训练得到的光场字典去恢复出待采集的光场。基于压缩感知的光场相机同时具有小体积和分辨率不损失的优点,但需要改造相机(在CCD表面插入一块编码过滤片),且其算法复杂度高、运算量大,目前还难以推向消费市场。

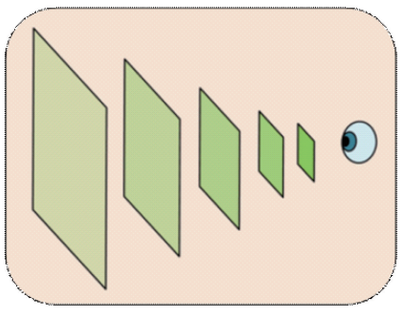

图12 支持多层聚焦成像的光场显示

Fig. 12 A light field display with multi-layer focus

6 VR与AR/MR

虚拟现实提供了强烈的沉浸感。佩戴者借助VR头显"穿越"到了一个完全由虚拟元素构成的世界中,但同时也把佩戴者与现实世界隔离开。在Virtual Reality的基础上,Augmented Reality(AR)应运而生。按照实现的技术方式,AR分为三类,包括Video-based AR,Optical-based AR和Projection-based AR。这三类AR都能实现真实场景和虚拟信息同时被人眼看见的视觉效果,但技术手段不同。

Video-based AR是对图片(或图片序列构成的视频)进行处理,在图片中添加虚拟信息,以帮助观看者进行分析和获得更多的信息。如图13所示,在手腕上添加不同款式的虚拟手表帮助消费者挑选合适的手表。再如时下热门的Faceu手机App,能在手机拍摄的图中添加诸如兔耳朵等可爱的虚拟元素。Video-based AR不需要佩戴特殊的眼镜,与观看传统平面图片方式一致,且允许非实时离线完成。

Optical-based AR通过类似半透半反的介质使人眼同时

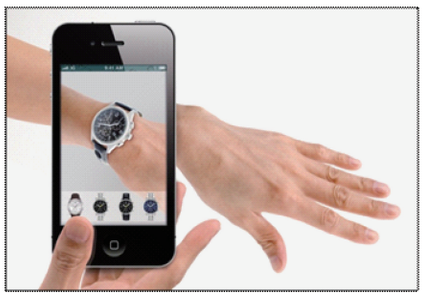

图13　基于Video-based AR的手表试戴

Fig. 13　A trial watch wearing based on video-based AR

（图片来源：http://www.cyingcg.com/article.asp?id=72）

接收来自真实场景和像源的光线，从而使得人眼同时看见真实场景和虚拟信息。Optical-based AR给人一种虚拟物体仿佛就位于真实场景中的视觉体验，但真实的场景中并不存在所看见的虚拟物体。且只有佩戴特殊头显设备（如Hololens，Meta）的人才能看见虚拟物体，没有佩戴头显设备的人不能看见虚拟物体。如图14所示，火箭模型并非真正存在于桌面上，且未带头显设备的人不能看见火箭。Optical-based AR相比于Video-based AR技术难度更大，需要三维环境感知。且从环境感知到增强显示都需要实时完成。

在虚拟现实行业出现了一个“新”的概念——MR（mixed

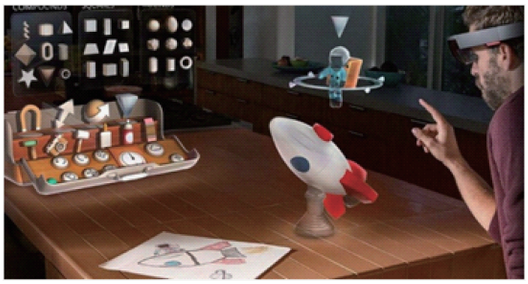

图14　Optical-based AR概念图

Fig. 14　Principle of optical-based AR

（图片来源：微软Hololens宣传视频）

reality），但这其实就是上述的Optical-based AR。图15是本文作者在实验室通过MR眼镜拍摄的照片，通过MR眼镜能同时看见真实的场景和虚拟的汽车。

Projection-based AR将虚拟信息直接投影到真实场景中物体的表面或等效的光路上。相比于Optical-based AR，Projection-based AR不需要佩戴头显设备却能获得与之类似的增强现实效果，且允许多人在一定角度范围内同时观看。图16为本文作者拍摄的基于投影增强现实的车载导航仪。路基线、车速、天气、来电等信息被投影在司机观看路面的等效光路上，司机不需要佩戴头显设备即可看见上述辅助信息。

虚拟现实带来了强烈的沉浸感但也隔断了人与人之间

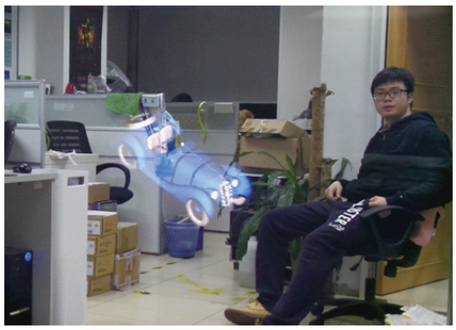

图15　混合虚拟现实——悬浮的小车（戴上眼镜后观看效果）

Fig. 15　Combined VR—a suspended car (the effects seen with glasses)

图16　基于Projection-based AR的车载导航

不佩戴眼镜观看效果

Fig. 16　Vehicle navigation based on projection-based AR

(the effects seen without glasses)

的联系。虽然人与人可以在虚拟世界中产生交互，但其交互手段有限，且交互的真实性和自然性都大打折扣。纵观历史上任何技术得以大面积普及的关键都在于密切的联系（dense communication）。从早期的互联网到智能手机及当前的移动互联网，之所以迅猛发展都离不开大量人群之间的通信。如果失去了人与人之间的通信也就失去成为大平台的基础。虚拟现实的隔断性注定了VR不会成为下一个智能手机。而MR弥补了VR的这一重大缺陷，能同时具备视觉信息增强和人人通信这两大特点。MR比VR有更高的概率成为智能手机在未来的新形态。

参考文献（References）

[1] Geng J. Three-dimensional display technologies[J]. Advances in Optics and Photonics, 2013, 5(4): 456-535.

[2] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-Time Single Camera SLAM[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(6): 1052-1067.

[3] Mackenzie K J, Watt S J. Eliminating accommodation-convergence conflicts in stereoscopic displays: Can multiple-focal-plane displays elicit continuous and consistent vergence and accommodation responses?[C]// Stereoscopic Displays and Applications XXI. Bellingham WA: SPIE,

2010. Doi: 10.1117/12.840283.
[4] Vienne C, Sorin L, Blondé L, et al. Effect of the accommodation-vergence conflict on vergence eye movements[J]. Vision Research, 2014, 100: 124-133.
[5] Hoffman D M, Banks M S. Disparity scaling in the presence of accommodation-vergence conflict[J]. Journal of Vision, 2010, 7(9): 824.
[6] Takaki Y. Generation of natural three-dimensional image by directional display: Solving accommodation- vergence conflict[J]. Ieice Technical Report Electronic Information Displays, 2006, 106: 21-26.
[7] Gershun A. The light field[J]. Mathematical Physics, 1939, 18: 51-151.
[8] Geng J. Volumetric 3D display for radiation therapy planning[J]. Journal of Display Technology, 2009, 4(4): 437-450.
[9] Van Berkel C. Image Preparation for 3D-LCD[C]//Stereoscopic Displays and Virtual Reality Systems VI. Bellingham WA: SPIE,1999. Doi: 10.1117/12.349368.
[10] Zhang Z X, Geng Z, Zhang M, et al. An interactive multiview 3D display system[C]// Emerging Digital Micromirror Device Based Systems and Applications V. Bellingham WA: SPIE, 2013. doi: 10.1117/12.2000360.
[11] Levoy M, Hanrahan P. Light field rendering[C]//Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 1996: 31-42.
[12] Jones A, Mcdowall I, Yamada H, et al. Rendering for an interactive 360° light field display[J]. Acm Transactions on Graphics, 2007, 26 (3): 40.
[13] Cao X, Geng Z, Zhang M, et al. Load-balancing multi-LCD light field display[C]//Stereoscopic Displays and Applications XXVI, 93910F. Bellingham WA: SPIE, 2015. Doi:10.1117/12.2078366.
[14] Cao X, Geng Z, Li T, et al. Accelerating decomposition of light field video for compressive multi-layer display[J]. Optics Express, 2015, 23 (26): 34007-34022.
[15] Bennett W, Neel J, Vaibhav V, et al. High-speed videography using a dense camera array[C]// Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2004: 294-301.
[16] Wilburn B, Joshi N, Vaish V, et al. High performance imaging using large camera arrays[J]. ACM Transactions on Graphics, 2005, 24(3): 765-776.
[17] Ren Ng, Levoy M, Bredif M, et al. Light field photography with a hand-held plenopic camera[R]. Palo Alto, California: Stanford University, 2005: Stanford Tech Report CTSR 2005-02.
[18] Marwah K, Wetzstein G, Bando Y, et al. Compressive light field photography using overcomplete dictionaries and optimized projections[J]. ACM Transactions on Graphics, 2013, 32(4): 46.
[19] Cao X, Geng Z, Li T. Dictionary-based light field acquisition using sparse camera array[J]. Optics Express, 2014, 22(20): 24081-24095.

Technological bottleneck of virtual reality

CAO Xuan[1,2]

1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. Institute for Creative Technology, University of Southern California, Los Angeles, CA 12015, USA

Abstract With the boost of both cutting-edge technologies and capital strength, the Virtual Reality(VR) makes a fast progress in recent years. And a technical maturity is reached for VR to be commercialized. With both the virtual reality and the 3D cinema, a stereoscopic display is generated based on the binocular parallax. But virtual reality provides more depth cues (e.g. the motion parallax) with a strong immersion than the 3D cinema. The current generation of the virtual reality still faces a series of technical problems. Especially, the vertigo and the eye fatigue are the toughest problems impeding the long-term applications of the VR. This paper briefly discusses the 3D visual perception and the causes of the vertigo and the eye fatigue. A solution to this technological bottleneck is suggested: the Dynamic Light Field. In addition, various light field technologies are compared related to the photograph and the display. At the end, three types of the augmented reality (AR) are presented and the commercial development trend of the AR and the VR are analyzed from perspectives of social interaction and communication.

Keywords virtual reality; augmented reality; dynamic light field; computational photograph; digital holographic display

（责任编辑 刘志远）