



专题：视频技术的理论与实践

虚拟现实视频处理与传输技术

董振江, 张东卓, 黄成, 刘海军
(中兴通讯股份有限公司, 江苏 南京 210012)

摘要: 虚拟现实(VR)是当前视频领域研究热点。介绍了VR视频内容制作以及终端显示技术现状,分析了基于用户视点的VR视频处理关键技术,最后探讨了VR视频传输模式及其对承载网络的需求。

关键词: 虚拟现实; 虚拟现实视频; 投影; 区域封装; 基于视点视频编码; 全景视频传输

中图分类号: TN919.8

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017244

Technology of virtual reality video processing and streaming

DONG Zhenjiang, ZHANG Dongzhuo, HUANG Cheng, LIU Haijun
ZTE Corporation, Nanjing 210012, China

Abstract: Virtual reality is a hot topic in the field of video. The current situation of VR video content production and VR video display technology was introduced, and the key technologies of viewport dependent VR video processing were analyzed. Finally, the mechanisms of VR video streaming and corresponding network requirements were discussed.

Key words: virtual reality, virtual reality video, projection, region-wise packing, viewport dependent video coding, omnidirectional video streaming

1 引言

虚拟现实(virtual reality, VR)因其沉浸感(immersion)、交互性(interaction)和构想性(imagination)为人们带来了更好的视觉体验。而VR视频需要经过内容制作、编码压缩、网络传输、终端显示等多个环节。本文将对这些环节进行介绍,并分析基于用户视点的VR视频处理与传输关键技术,最后探讨VR视频业务对承载网络的需求。

2 VR视频内容

VR内容形态上包括VR视频和VR CG(computer graphic, 计算机图形)两大类,前者主要是通过摄像机拍摄并生成全景(panorama)视频(单目或双目),后者主要是通过计算机图形生成三维模型。本文主要讨论前者。

2.1 VR视频制作

VR视频生成的方式有两种:一种是用摄像机拍摄;另一种是采用CG技术3D建模生成视频。



摄像机拍摄 VR 视频, 需要用全景摄像机即多镜头摄像机拍摄各个方向的图像内容并进行图像拼接。目前市场上全景摄像机的镜头(含镜头后面的成像设备等)数量从两个到十几个不等。如果要生成双目全景视频, 则每个方向上至少覆盖两个镜头。而有些全景摄像机还配置了激光雷达等深度测量设备。要生成分辨率更高的 VR 视频, 则需要更高分辨率的摄像机或者更多摄像机镜头。如何降低高分辨率全景相机的成本也是未来 VR 高质量视频内容进行市场普及的一个重点问题。由 VR CG 生成的视频则类似于 3D 动画的 VR 版本, 与全景拍摄相比, 其最大的优势是节奏控制和调度方便, 尤其是辅以各种贴图、渲染和光效, 其视觉效果经常可以以假乱真。在未来, VR 视频内容如六自由度视频和自由视点视频, 可能需要摄像机拍摄和 CG 建模两种技术手段的融合使用。

2.1.1 三自由度 VR 视频

(1) 单目全景视频

拍摄和拼合出来的画面是一个球面, 球面电影就是一种单目全景(只是缺少下面一小部分), 目前在网络中多数 VR 视频是单目全景, 即左右双眼看到的内容是一致的。这种全景视频的一个缺陷就是没有立体感, 对事物的距离感体验有缺失。

(2) 双目全景视频

双目全景视频利用双目视差原理, 在 VR 头盔的左右双眼中呈现具有一定视角差的影像, 原理上与 3D 电影别无二致。只是在双目全景视频的生成过程中, 可以通过并列双镜头摄像机拍摄生成, 也可以通过具有深度信息的单镜头摄像机拍摄并通过深度计算呈现左右眼视差。

(3) VR180 视频

日前, YouTube 和微软联合推出了一种的全新的视频格式, 名为 VR180, 其相当于全景视频的前面一半, 且全景的顶部和底部也进行了消减,

可在 VR HMD 或普通电视/PC 屏幕或移动端收看。VR180 只保留最核心局域, 尤其适合观看现场直播类的 VR 内容如赛场、演唱会、会议、谈话节目等。但是 VR180 无法替代需要更强沉浸感的全景/VR 视频。VR180 可降低拍摄成本, 拍摄简便, 减少侵权风险, 后期或即时制作更便捷, 降低 VR 内容门槛, 促进 UGC 产生。同时, VR180 同等分辨率下要求网络带宽低, 终端计算能力低, 终端可获得性高, 促进了大屏 VR 业务发展。

总体来看, VR180 是现阶段技术水平限制性的明智务实的折中, 将促进 VR 内容产生。

2.1.2 六自由度 VR 视频

自由度(degrees-of-freedom, DoF)表示物体的可移动维度, 目前多数 VR HMD 和 VR 视频只支持三自由度, 即在特定观察位置, 实现头部方位角旋转(yaw)、俯仰角旋转(pitch)和摇摆旋转(roll)。六自由度是在三自由度的基础上, 加上观看位置的空间变动: X 、 Y 、 Z 。

六自由度视频生成有如下 3 种方式: 光场相机记录、3D 模型、空间捕捉。

六自由度如果用光场相机(多个相机组成的矩阵)记录, 则流量暴增(对多角度的视频压缩也是一个巨大的挑战), 且自由观看的移动范围有限。3D 模型中的六自由度, 类似于 3D 游戏中主观视点和视角的变动, 实际上属于 VR CG 范围而不是 VR 视频范围, 也实现了自由视点。所谓空间捕捉方式, 是前两种手段的一种融合。在摄像机拍摄同时通过激光雷达扫描生成四周环境的 3D 模型, 之后进行贴图。可以一定程度上更方便快捷地生成六自由度视频, 但它生成的环境模型基于一个视点, 所以 3D 环境模型和物体模型有一定的偏差。

尽管六自由度还处于起步阶段, 但 VR 发展到六自由度是更高沉浸感的未来必然方向。

2.2 VR 视频显示

对于单目全景视频还有平面显示、环幕显示和球幕显示方式, 但 VR 视频显示的绝对主力是

VR 头盔 (head mounted display, HMD)。总体看来, VR 视频显示方面需要优化演进的包括如下几个领域: 视听效果, 包括显示分辨率、视场角等; VR HMD 的显示计算能力; 更好的佩戴舒适性和移动性。

2.2.1 显示分辨率

目前 VR HMD 最主要的短板在于显示分辨率。VR HMD 显示分辨率的更准确定义不是像素数, 而是单位角度的像素数 (pixel per degree, PPD)。PPD 的定义是每度视角有多少个像素。根据人眼的识别能力, 视力 1.0 对应的 PPD 是 60, 即每度视角里有 60 个像素点, 能达到 60PPD 的 VR HMD 将是一个比较理想的显示分辨率。而目前主流 VR HMD 产品的 PPD 参数在 10~15 内, 对应的头盔屏幕分辨率为 2K 左右, 距离理想效果的要求还有较大的差距。目前已经有几款 4K 分辨率的头盔上市, 预计 2018 年将会进入 20PPD 的阶段。预计 VR 头盔显示分辨率的演进见表 1。

2.2.2 视场角

FOV (field of view) 角度是 VR HMD 与平面视频不同的另一个重要参数, 它表示了带上 VR HMD 后, 人眼可以看到 VR 内容的范围。人的双眼在水平范围的 FOV 可以达到 200°, 而目前主流商用 VR HMD 的 FOV 为 96°~120°, 距离完全沉浸还有较大的距离。当然人眼在水平 60°范围之外的视觉敏感性很低, 这些范围并不需要很高的显示分辨率, 可以通过光学处理, 用较少的显示像素覆盖较大的人眼余光视角范围, 在提升沉浸感

的同时提高显示效率。

2.2.3 VR HMD 的显示计算能力

VR 视频显示需要 VR HMD 对压缩视频进行解压、球面投影、渲染等工作, 需要一定的显示计算力。即需要 GPU (用于外接式 VR 头盔) 或者相应的处理器 (VR 一体机或手机式 VR 头盔)。而目前 VR 头盔分辨率显示的瓶颈就是 VR HMD 的显示计算力。尤其是随着 VR 视频分辨率的提升, 图像显示计算量呈指数增长, 这一方面需要集成电路芯片领域持续地提升计算力降低单位计算能力成本, 另一方面也需要 VR 领域图形算法的持续优化, 以降低 VR 图像生成的计算量。

VR 视频比起 VR CG, 在显示计算力方面的需求较小, 因此 VR 一体机和移动式 VR 头盔也会有较好的显示效果。VR 一体机在成本和效果之间有一个较好的平衡, 未来预计 VR 一体机将是 VR 视频的主流观看平台。

2.3 VR 视频体验

从用户体验来讲, VR 视频内容与平面视频相比, 有更强的沉浸性、交互性和构想性。但 VR 视频由于节奏、调度、镜头语言与平面视频有很大差别, 因此 VR 视频的题材范围差异也很大。VR 视频并不适合所有的题材, 其适合如下内容: VR 体育直播/大型活动直播, VR 成人视频点播/直播, VR 动画影视 (相当于 VR 版本的动画片), VR 旅游/房产视频点播/直播, VR 综艺节目、谈话节目、真人秀, 对话推进的 VR 情景影视, UGC VR 游戏/漫游转录的 VR 视频点播/直播。

表 1 VR HMD 商用阶段现状和预测

阶段 (显示屏分辨率)	阶段 1 (2K VR)	阶段 2 (4K VR)	阶段 3 (10K VR)	阶段 4 (极致体验 VR)
商用时间段	2016—	2018—	2020—	2022—
头盔显示屏分辨率	2 160 dpi×1 200 dpi	4 320 dpi×2 400 dpi	9 600 dpi×4 800 dpi	14 400 dpi×7 200 dpi
单眼视角内像素/帧频	1 080 dpi×1 200 dpi/60 帧/s	2 160 dpi×2 400 dpi/90 帧/s	4 800 dpi×4 800 dpi/120 帧/s	7 200 dpi×7 200 dpi/120 帧/s
单眼视角 FOV	110°×120°	110°×120°	120°×120°	120°×120°
像素密度	10	20	40	60

头和传感器的摄像设备以及音频传感器录制而成。设备的输出是一组数字视频和音频信号。摄像头通常可以获取设备中心周围所有方向的内容,故名全景视频或者 360°视频。

其中,音频可以通过不同的麦克风配置来获取,可以用不同格式存储,包括基于通道信号、静态或动态(在 3D 场景中移动)对象信号以及基于场景的信号。在一个虚拟现实应用中,扬声器阵列的信号通过沉浸式音频渲染程序处理后被量化,然后通过耳机呈现给用户。

对于虚拟现实视频而言,同一时刻的 VR 视频图像经过拼接、投影映射成一个封装帧,如图 2 所示。

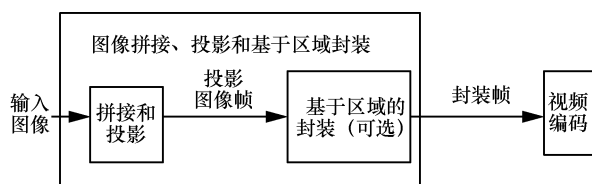


图 2 虚拟现实图像拼接、映射和基于区域封装

与传统视频相比,虚拟现实视频的独特特征在于通常仅显示对应于当前视点(即用户当前观看区域)的全景视频中的部分视频。可以利用该特征,采用视点相关虚拟现实视频处理技术,包括:视点相关投影映射技术、基于视点的区域封装技术以及基于视点的视频编码技术,在保证为用户提供相同视频分辨率/质量的条件下,减少视频传输带宽需求,或者降低视频解码复杂度,从而改善虚拟现实视频系统整体性能。

3.1 投影映射

输入图像经过拼接后投影到一个三维投影结构上,例如一个单位球体或者正六面方体。投影结构上的图像数据将被进一步地排布到一个二维平面的投影帧^[1]。

ERP (equirectangular projection, 等角投影)是目前 VR 视频主流投影格式,如图 3 所示。

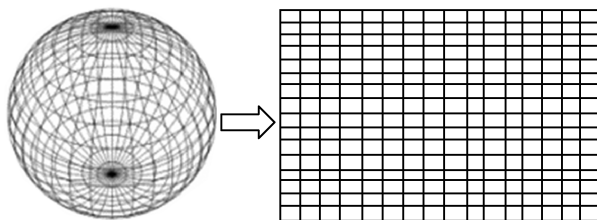


图 3 ERP 投影格式

ERP 投影类似于地球球面展开成地图,将球面展开为平面矩形,所以也被称为经纬图展开。这种方式简单并已经普遍使用,但其缺点也是明显的:球面赤道部分投影展开后失真小,而两极部分纬度越高,失真越大。同时引入了过多的无效像素,球面展开成 ERP 后,面积即像素数增加了 57%,从而降低了传输效率。

PSP (platonic solid projection, 正多面体投影)是业界关注的新方向,具有失真小、压缩效率高的特点。正多面体投影将球体以球心为中心向外划分为多个球面区域,并投影到多面体某个面上,正多面体可以是四面体、立方体、金字塔、十二面体和二十面体等,具体见表 2。由于每个球面区域单独投影,因此失真较小。对于面数多的方案,其总面积增加也较小,例如正二十面体投影方案的总面积只有球面面积的 1.21 倍。同时,正多面体投影格式有利于实现基于视点 VR 视频传输方案,传输部分“面”的视频图像,或者对于不同“面”提供不同的分辨率,这样可以降低传输带宽需求。

表 2 正多面体投影格式

投影结构	三维模型	二维映射	顶点	面积比率
正四面体			4	3.31× 2.11×
正六面体			8	1.91× 1.22×
正八面体			6	1.65× 1.05×
正十二面体			20	1.32× 0.84×
正二十面体			12	1.21× 0.77×



对于单目虚拟现实视频,同一时刻的输入图像拼接后生成一个投影帧,用来代表一个视场。对于双目立体虚拟现实视频,同一时刻的输入图像拼接后生成的一个投影帧代表两个视场:左眼和右眼。两个视场可以映射到相同的封装帧,并基于传统 2D 视频编码器进行编码。或者投影数据帧的不同视场可以被映射到各自的封装帧。左视场或者右视场的封装帧序列可以被单独编码,或基于多视点编码技术进行左右视场预测编码。

3.2 区域封装

基于区域封装是对投影后形成的投影帧进行重新映射,生成一个或多个封装帧的过程。投影帧中不同的投影区域按照不同的采样率以及旋转(90° 、 180° 、 270°)、镜像(水平)等转换操作,被映射至封装帧中对应的封装区域。例如, CMP (cube map projection, 正六面体投影) 中可以针对立方体某个面按照更高的分辨率进行采样,并将 6 个面映射到同一封装帧,如图 4 所示。



图 4 基于区域封装示例

针对同一内容源,可以多次执行上述虚拟现实图像拼接、投影和基于区域封装处理,以创建相同虚拟现实内容的不同版本。类似地,针对相同的投影图像帧多次执行基于区域封装处理,可以创建用于编码的多个封装帧图像序列。

3.3 基于视点编码

不同于传统视频编码方案,VR 视频内容在编码之前可以被分割成子图像序列或者运动约束瓦片集合。以前者为例,每个子图像序列覆盖全景视频内容空间区域的子集,彼此独立地编码为单层比特流,其中,同一子图像序列可被编码为多个不同码率的比特流。

每个子图像比特流作为独立的轨道被封装在一个视频文件中用于流式传输。接收方基于视点朝向等元数据选择要传输的视频轨道,接收方通常会接收覆盖全方位区域的视频轨道,其中对应当前视点区域的视频轨道质量或分辨率应高于不可见区域的视频轨道。

每个子图像比特率必须使用独立解码器进行解码。为了减少用于解码对应当前视点子图像比特流所需视频解码器的数量,VR 视频中划分的子图像之间可以具有重叠区域。如图 5 所示,子图像 1 由区域 A 和 B 组成,子图像 2 由 B 区和 C 区组成,子图像 3 由 C 和 D 区组成,子图像 4 由 D 区和 E 区组成。子图像 5 由 E 和 A 区组成,子图像 6 由区域 F 组成,子图像 7 由区域 G 组成。

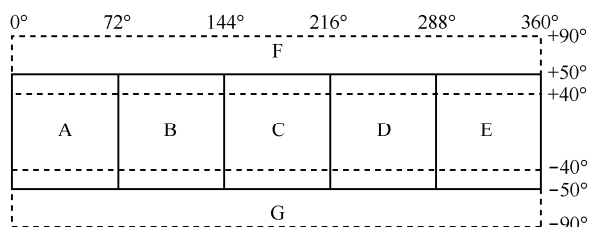


图 5 基于视点图像编码示例

该方案中使用 1 个子图像即可渲染出靠近赤道的任何视点区域,当用户视点靠近南北极时最多再需要解码 1 个子图像,从而可以预先知道解码当前视点所需的最多解码器

4 全景视频传输

4.1 传输模式

FOV 是虚拟现实视频或者全景视频(omnidirectional video)的基本属性,虚拟现实视频可以具有预先确定的主视场(primary FOV),主视场内的视频质量(分辨率)高于周围视频质量。此外,虚拟现实视频还可以划分成为具有预先确定辅助视场(supplemental FOV)的分块视频,低分辨率的虚拟现实全景视频结合高分辨率的分块视频,可以在预先确定的视场范围内获得较高的视频质量^[2]。

(1) 视点独立全景视频传输

在视点独立全景视频传输模式（简称为基本传输模式）下，虚拟现实视频中不存在预先确定的主视场。客户端向服务器请求获取无差别的全景视频文件，当用户视角发生变化时，所有的处理都在终端完成。

(2) 视点自适应全景视频传输

在主视场视点自适应传输模式（简称为 FOV 传输模式）下，虚拟现实视频中存在预先确定的主视场。当用户视角变化量大于特定阈值时，客户端确定当前用户视角并根据用户视角与视频文件的对应关系，向服务器请求获取对应于用户当前视点的全景视频文件。

在辅助场视点自适应传输模式下（简称为混合传输模式），虚拟现实视频中存在预先确定的辅

助视场。客户端向服务器请求获取无差别的全景视频文件，当用户视角变化量大于特定阈值时，客户端确定当前用户视角并根据用户视角与视频文件的对应关系，向服务器请求获取对应于用户当前视点的辅助视频文件。

类似地，在分块传输模式下，客户端确定当前用户视角并根据用户视角与视频文件的对应关系，向服务器请求获取对应于用户当前视点的分块视频文件。

4.2 网络需求

根据各种 VR 视频类型的特点，可以定性分析出其对网络的需求，包括网络存储分发、网络计算需求、网络带宽需求、网络时延需求等，具体见表 3。

对于带宽需求最高的全景视频，可以定性估算一下其各个阶段具体的网络带宽和时延需求，见表 4。

表 3 各种类型 VR 视频对网络的需求

VR 视频类型	视频格式	本地计算需求	网络存储分发需求	网络计算需求	网络带宽需求	网络时延需求
本地 VR 视频	全景	高	无	无	下载	无
全景 VR 网络视频	全景	中低	标准	无	高	低
预制 FOV VR 视频	预存 FOV	低	高	低	中	中
即时 FOV VR 视频 (网络即时计算生成 FOV 视频)	即时 FOV	低	标准	高	中	中高

表 4 VR 视频商用阶段现状和网络需求预测

阶段	阶段 1 (2K)	阶段 2 (4K)	阶段 3 (10K)	阶段 4 (极致体验 VR)
商用时间段	2016—	2018—	2020—	2022—
头盔显示屏分辨率	2 160 dpi×1 200 dpi	4 320 dpi×2 400 dpi	9 600 dpi×4 800 dpi	14 400 dpi×7 200 dpi
单眼视角内像素/频率	1 080 dpi×1 200 dpi/60 帧/s	2 160 dpi×2 400 dpi/90 帧/s	4 800 dpi×4 800 dpi/120 帧/s	7 200 dpi×7 200 dpi/120 帧/s
单眼视角 FOV	110°×120°	110°×120°	120°×120°	120°×120°
像素密度	10	20	40	60
全景总像素/频率/色深 (注 1)	3.98×10 ⁶ dpi/60 帧/s/8 bit	2×15.9×10 ⁶ dpi/90 帧/s/10 bit	2×66×10 ⁶ dpi/120 帧/s/12 bit	2×148.5×10 ⁶ dpi/120 帧/s/12 bit
VR 视频投影格式 (注 2)	ERP (×1.57)	ERP (×1.57)	ISP (×1.21)	ISP (×1.21)
原始无压缩码率/ (Gbit·s ⁻¹)	3	45	230	517
压缩比	200(H.265)	200(H.265)	400(H.266)	400(H.266)
压缩码率/ (Mbit·s ⁻¹)	15	225	575	1 294
带宽需求/ (Mbit·s ⁻¹)	22	337	863	1 940
传输往返时延 RTT/ms	20	16	12	10
分组丢失率 (TCP 模式)	7.05×10 ⁻⁴	4.69×10 ⁻⁶	1.27×10 ⁻⁶	3.62×10 ⁻⁷
分组丢失率 (UDP 模式)	6.34×10 ⁻⁷	5.63×10 ⁻⁷	5.58×10 ⁻⁷	5.58×10 ⁻⁷

注 1: 阶段 1 为单目视频，阶段 2 及以后为双目视频。

注 2: 括号里的数字表示此格式与对应球表面积的比例。



可见商用阶段 VR 视频对网络带宽需求非常高。这里如果采用视点自适应视频传输模式，其带宽需求将大幅度降低（具体比例由具体传输模式的设定会有较大不同）。而多面体投影模式的不同面给予不同分辨率也同样可以降低带宽需求。当然这需要采集用户的视角方向信息，和全景传输模式相比，对网络时延有更高的要求。

5 结束语

本文对 VR 视频内容制作、终端显示环节技术现状进行了介绍，分析了基于用户视点的 VR 视频处理关键技术，最后探讨 VR 视频传输模式及其对承载网络的需求。

随着各方面支撑技术的持续发展，预期 VR 视频将不断向着更高的分辨率、更广的视场角、更高效的编码压缩与实时传输方向发展演进。未来市场上将会出现越来越多更平民化的成本、更友好的沉浸体验、更丰富精彩的 VR 视频业务与应用，进而影响并丰富人们生产生活的各个方面。

参考文献：

- [1] YU M, LAKSHMAN H, GIROD B. A framework to evaluate omnidirectional video coding schemes[C]//2015 IEEE International Symposium on Mixed and Augmented Reality, September 29-October. 3, 2015, Fukuoka, Japan. New Jersey: IEEE Press, 2015: 31-36.
- [2] CORBILLON X, DEVLIC A, SIMON G, et al. View-port-adaptive navigable 360-degree video delivery[C]//ACM

Multimedia Systems 2016 & NOSSDAV, MoVid, and MMVE, May 10-13, 2016, Klagenfurt, Austria. New York: ACM Press, 2016.

[作者简介]



董振江（1970—），男，中兴通讯股份有限公司研究员级高级工程师、云计算及 IT 研究院副院长，中国人工智能学会常务理事，中国计算机学会服务计算常务委员，主要研究方向为人工智能、AR/VR、媒体分析与处理等。



张东卓（1974—），中兴通讯股份有限公司资深架构师，主要研究方向为虚拟现实和增强现实。



黄成（1978—），男，中兴通讯股份有限公司高级系统工程师，主要研究方向为视频传输与多媒体通信系统。



刘海军（1970—），男，中兴通讯股份有限公司预研项目经理，主要研究方向为智能家居业务与技术，包括视频业务、人机交互技术、人工智能应用。