

基于多视角视频的目标识别与虚拟化身生成系统

文/纪红蕾 沈晓宇
中车工业研究院有限公司

摘要：虚拟现实技术在交通领域具有广泛的应用前景，可以提升工作效率、降低成本。该系统利用虚拟现实技术，基于多视角视频的目标识别与虚拟化身生成技术构建地铁驾驶舱虚拟孪生系统，实现了对真实驾驶舱的快速准确监控，全方位提升监控系统的感知能力。该地铁驾驶舱虚拟孪生系统为构建集感知、调度、管理、应急及服务于一体的地铁数字孪生平台奠定了基础。

关键词：虚拟现实技术；三维虚拟驾驶舱；多视角视频；目标检测；沉浸式体验

虚拟现实技术打破了人们在地理空间上的限制，使数字交通以更直观的方式展示和更高效的方式管理，并以更智能的方式进行处理。同时，虚实结合的操作环境能够提升工作效率，降低产品开发成本。在交通运输领域，三维仿真技术可有效模拟各种场景，为复杂条件下的交通提供工具技术支持。此外，虚拟现实技术还可应用于驾驶员的培训和各种测试，突破了地域限制，实现多人协同工作，降低了人力资源成本。在产品设计和开发方面，利用虚拟现实技术构建的仿真平台可用于交通设计，并分析产品的安全性能、使用体验等，降低了产品的生产周期与成本。因此，虚拟现实技术在交通领域具有广泛的应用前景。

本系统将虚拟现实技术应用于轨道交通，通过基于多视角视频的目标识别与虚拟化身生成技术构建地铁驾驶舱虚拟孪生系统，旨在全方位提升监控系统的感知能力。构建与地铁驾驶舱主体物理环境一致的三维虚拟驾驶舱，是本系统将虚拟现实技术应用于轨道交通的关键之一。本系统基于驾驶操纵台CAD图和顶部、控制台的两个同步监控视频序列，能够识别

司机驾驶行为、操作台显示灯等，并重现驾驶员的三维虚拟化身，支持通过虚拟驾驶舱对真实驾驶舱的自由视角监控。在技术实现方面，本系统在有限视角下能快速准确识别人物姿态、手势动作和交互物体。同时，通过接收二维视频内容，项目所用模型能够准确地定位人物的三维空间坐标。

本系统的研究主要分为三个方面：首先，研究多视角视频三维姿态识别技术，以实现基于顶部和控制台同步视频的驾驶员姿态识别与三维重建，具体来说，本系统从输入的视频内容进行驾驶员的姿态识别，然后进行三维姿态拟合；其次，研究目标物体检测技术，以实现驾驶员交互物体的检测识别，重点关注手部区域与可交互物体的检测，从而实现更加精准的交互体验；最后，研究基于Unity引擎开发的虚拟驾驶舱系统，该系统可以在PC端进行浏览，并且可以在VR端达到沉浸式体验的效果，通过该地铁驾驶舱虚拟孪生系统，用户可以实现对地铁司机驾驶室的沉浸式全方位监控，为构建一个集感知、调度、管理、应急及服务于一体的地铁数字孪生平台奠定基础。

相关工作

近年来,人体重建技术在计算机视觉领域获得了广泛关注和应用,一些学者提出了具有代表性的算法。Neural Body是一种基于NeRF和可变人体模型(如SMPL)相结合的方法,可对视频中的人体进行三维网格重建和新视角合成。这种方法在处理不同外观或动作的人物时需要重新训练,而PixelNeRF等通用场景的可泛化NeRF算法难以处理身体部位相互遮挡的问题。为了解决这些问题,Kwon等人提出了Neural Human Performer,该方法使用Transformer聚合多视角视频的时序特征和多视角特征,在多个不同场景上对模型进行预训练,之后只需要少量新数据对模型进行微调,即可获得比从头训练更优的新场景渲染效果。Animatable-NeRF利用NeRF从多视角视频中重建了人体模型,并引入了神经混合权重场,可以与特定动作的骨骼相结合,生成新的变形场,渲染出利用骨架驱动的人体视频和三维网格模型。此外,Liu等人提出的Neural Actor可以在渲染阶段调节渲染结果中人的体型,而HumanNeRF可以处理人类舞蹈等复杂运动的单目视频。

人脸重建方面,Gafni等人提出了4D-Facial-Avatars,其结合NeRF和人脸模型,得到静态外观和动态姿态、表情信息,渲染得到更多不同表情、姿态的人脸图像。Guo等人提出了AD-NeRF,利用视频的声音和图像信息训练NeRF模型,渲染阶段输入新音频信息,得到音频驱动的新视角视频。Hong等人提出了HeadNeRF,将NeRF与人体头部参数相结合,直接控制人体头部的姿态和语义属性,同时将2D神经渲染集成到NeRF中,设计新损失函数,提

高渲染准确度和速度。Sun等人提出了FENeRF,将GAN与NeRF结合,实现语义级别的人脸编辑和新视角合成。

系统算法介绍

地铁驾驶舱虚拟孪生系统的核心算法主要包括六部分:①具备车外视野矫正功能,可以将车外视角内容映射到驾驶舱的玻璃面板上;②能够重建脸部姿态,通过正面视角内容对人脸的表情姿态进行进一步预测;③拥有人体姿态识别功能,可以直接检测出视频中人物的运动姿态;④能识别手势动作,对视频中人物的手部动作进行检测和识别;⑤能够检测交互物体,识别人物与操控台的交互物体,如单击按钮、推动滑杆及拿起对讲机等;⑥对预测的人体姿态进行矫正,依据识别的角色姿态以及交互物体检测结果,进一步对角色手部姿态、交互位置以及角色行走坐下姿态进行矫正与平滑。这些功能的共同作用可在虚拟地铁驾驶舱中准确重建驾驶员的操作与行为,实现虚拟孪生地铁驾驶室的精准构建。

1. 系统总体架构

该系统采用Unity三维引擎编写系统,利用驾驶舱采集3个不同视角视频,同步生成对应的虚拟内容。执行方案总体架构如图1所示。为用户提供新型浏览方式,用户可以通过VR设备(如头盔、手柄等)或直接利用鼠标、键盘自由漫游虚拟三维环境。系统提供前端用户交互系统、后端系统通信服务以及目标识别与姿态重建模型。相较于传统的观看视频的方式,该系统提供更详细的内容及更直观的视觉体验。

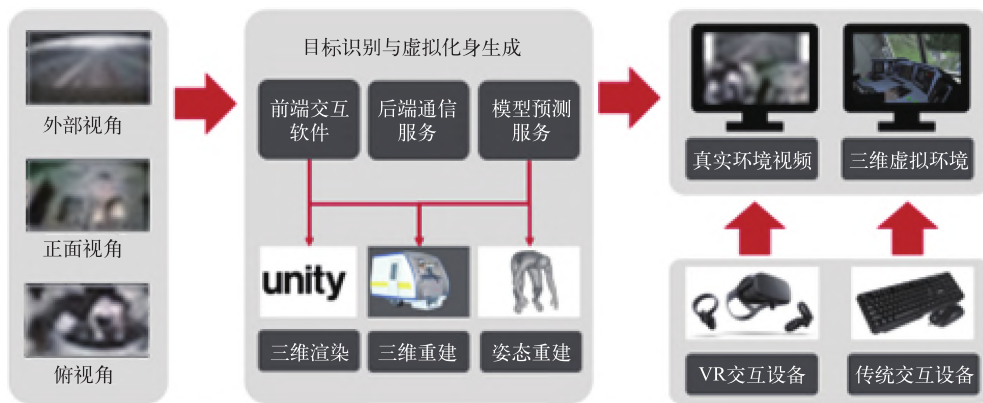


图1 执行方案总体架构

该系统实现了前端交互三维建模，包括驾驶舱建模和控制台建模。驾驶舱三维建模如图2所示。首先根据监控视频的车厢参考图像，利用Maya手工制作模型的三维网格，其次，对模型进行UV纹理映射以及贴图绘制，获得具有贴图纹理的车厢模型，控制台三维建模如图3所示，根据提供的二维CAD设计图样制作控制台三维模型元件；然后，根据车厢参考图对元件进行拼合，形成三维控制台模型；最后，根据车厢参考图绘制模型纹理贴图，为控制台模型绑定骨骼，以满足交互时动画需求。

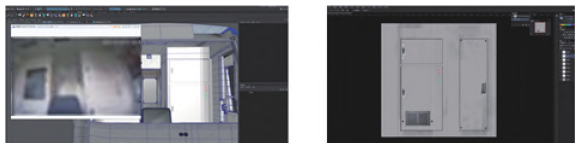


图2 驾驶舱三维建模



图3 控制台三维建模

为了实现用户与虚拟驾驶舱的人机交互，设计了前端交互系统，如图4所示。在系统UI中，主菜单包括文件、交互、偏好及帮助等模块，用户以文件夹浏览的方式导入监控视频，切换键盘/VR交互模式，调整用户偏好及获取说明帮助等功能。侧边栏提供真实视频参考，允许用户调整当前播放进度。系统将自动同步人物动画和视频播放进度，以使用户更好地进行实时交互。



图4 前端交互系统

为了实现前端交互与后端预测功能分离设计，采用Web通信方式实现前后端异步请求与处理。为了方便功能开发与问题维护并实现跨平台应用部署，采用Docker容器服务构建系统，如图5所示。Docker容器服务包含异步请求处理服务、消息存

储服务和深度模型处理服务。异步请求处理服务为用户的每次请求分配全局唯一的请求ID，并将状态变化信息存储到消息存储服务中，用户可以随时获取请求在后端处理的状态。深度模型处理服务依据请求提供的视频地址或视频流，应用深度网络模型预测视频中角色的姿态与交互物体。此外，整体架构设计允许前后端服务物理分离，用户可以向后端Docker容器网络地址发送处理请求，同时可以部署多个前端系统。

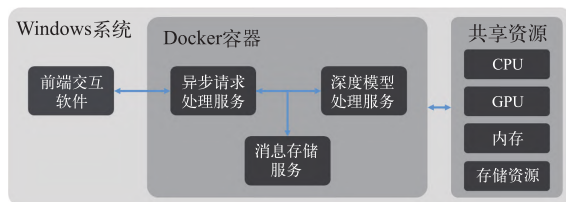


图5 应用总体结构图

2. 车外视野矫正

车外视野矫正的总流程如图6所示。车外视野矫正的总流程包括从相机视角中确定目标区域变换参数、获取原始视频中目标区域、提取视频每一帧对应的目标区域，并将目标区域纹理映射到驾驶舱玻璃模型。其中，提取目标区域的核心流程是：首先对原始图像进行三角网格化；然后手动标注目标区域位置，并分别将图像上下两行标记点投影到边缘；最后对所有三角网络进行变换和内部插值，最终输出目标区域。提取目标区域的流程如图7所示。

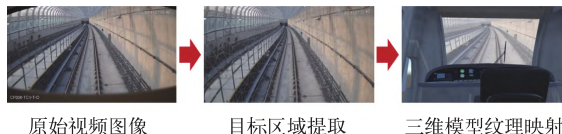


图6 车外视野矫正流程

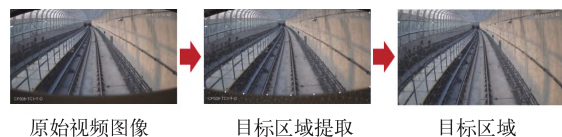


图7 目标区域提取流程图

为了解决纹理过小或过大时产生的失真走样问题，车外视野矫正采用了多级渐远纹理。具体来说，当与观察者的距离超过一定的阈值时，使用不同的多级渐远纹理，即最适合物体距离的那个。由

于这些距离较远，即使解析度不高，也不会被用户注意到。同时，由于不同距离采用不同的纹理，可以大大优化性能。这种处理方式可以有效地解决失真走样等纹理相关问题。

3. 重建脸部姿态

重建脸部姿态流程如图8所示。首先，采用现有的检测算法（例如Faster-RCNN）对人脸区域进行检测；然后，使用3DMM或FLAME模型预测人脸表情参数；最后，结合模型中的其他姿态参数，完成对人体姿态的重建。

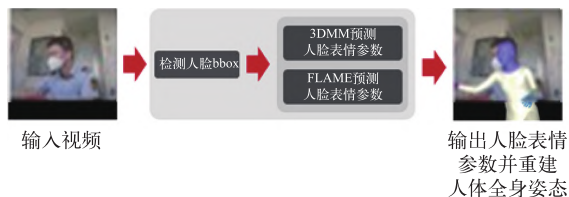


图8 重建脸部姿态流程

4. 人体姿态识别

人体姿态识别流程包括检测人体所处区域的包围盒（bbox），从包围盒中提取人体各个部位的2D骨骼点，并将这些点转换为3D骨骼点。然后，将3D骨骼点输入给人体模型，以得到最终的三维模型。流程如图9所示。

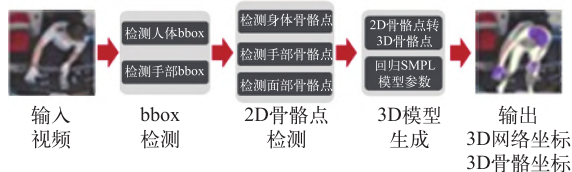


图9 人体姿态识别流程

为了检测人体所处区域的包围盒，利用了Fast-CNN和Cascade RCNN两种目标检测模型。模型的输入为单张RGB图片，输出为检测到的矩形区域的xy坐标。模型支持对一张图片里的多个目标同时进行检测，并输出每个目标的区域坐标。

系统选择了Heatmap的方法检测全身2D骨骼点，使用HRNet+Heatmap的方法在COCO-Wholebody数据集下进行训练，可从经过裁剪得到的人体图片中准确检测出人体的各个骨骼点。由于2D骨骼点不能对人体进行准确的建模，系统实现了从2D骨骼点到3D骨骼点的转换（2D-to-3D lifting），从而使用更准确的3D骨骼点建模。系统

使用在Human3.6M训练集下训练的lifting网络。该网络通过两层残差连接的全连接层，实现了2D骨骼点到3D骨骼点的一一映射。

利用SMPL（Skinned Multi-Person Linear）模型，系统实现了从身体3D骨骼点到可视化人体模型的转换。该模型需要输入3D骨骼点作为人体姿态特征，对原始的人体模型进行几何变换后得到该姿态下的人体模型。更进一步，通过结合手部和面部的模型，系统最终综合出了SMPL-X模型，用以表示全身模型。SMPL-X模型的训练基于PIXIE算法。PIXIE算法融合了脸部、手部和人体SMPL-X参数，最终合成得到的SMPL-X模型包含驾驶员的姿态信息和手部、脸部的细节信息，可以用Unity进行渲染。

5. 人体手势识别

手势识别流程如图10所示。经过全身姿态检测后，通过骨骼点的位置获得手部的包围盒。在此基础上，直接使用手部的2D坐标进行一些简单的检测，也可以使用其他网络获得3D的骨骼点。最后，使用SMPL-H网络预测出手部较为精细的模型，并与身体其他部位结合得到最终的人体模型。



图10 手势识别流程

6. 交互物体检测

视频角色姿态识别之后，还需判断角色是否与操作台产生交互动作以生成操作台对应的响应事件。依据现有监控视频内容以及操作台三维模型，系统对操作台上33个交互部件进行了标注，以确定交互部件数量、类型及每个部件的区域范围。如图11所示，每一个部件标注特定的编号，并确定其在俯视角中的图像位置。



图11 操作台交互部件区域示意图

7. 人体姿态矫正与平滑

为了解决驾驶员模型手部自由运动时可能出现的穿模现象,需要对其姿态进行矫正。该矫正方法从空间层面出发,包括水平方向和垂直方向上的约束。水平方向的约束通过将操作台水平面范围内的水平坐标进行线性映射来实现,从而保证驾驶员模型的手部运动和原始视频中的近似一致。垂直方向的约束限定了手部运动在一定高度范围内,防止手部与操作台面发生不合理的接触式穿模现象,并且使驾驶员模型能够快速响应特定的交互操作。在前端系统中,需要将后端获取的驾驶员手部运动姿态解析成被驱动人物模型的时空序列,并将手部运动限定在合理的区域内。这样可有效降低模型穿模风

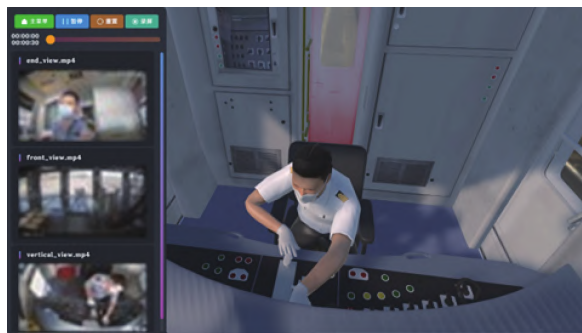
险,并提高整体视觉表现。

系统实现情况

系统运行所需基本硬件主要包括计算机主机与VR设备两部分,其中计算机主机CPU建议使用Intel® Core™ i5-4590 及 AMD Ryzen 1500 或更高版本;GPU显卡建议采用GeForce® RTX 20 Series (Turing) 及 AMD Radeon™ 5000 (Navi) 或更高版本;内存8 GB RAM或以上;视频输出需要带DSC的DisplayPort 1.4或更高版本。此外需要多个USB端口以连接各种外设设备(VR头盔所需),操作系统建议使用Windows 10或Windows 11。系统各部分示意图如图12所示。



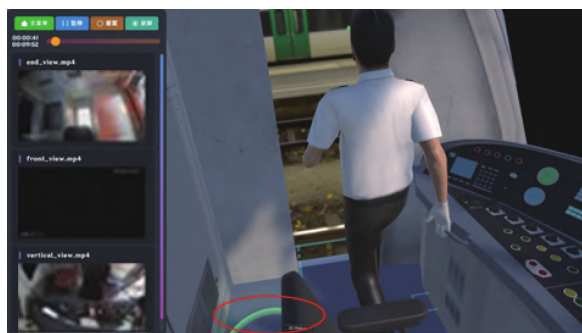
(a) 主界面视频加载展示图



(b) 视频解析虚拟动画展示图



(c) VR端使用交互模式设置图



(d) VR端用户利用手柄移动展示图

图12 系统各部分示意图

结束语

虚拟现实技术在交通运输领域具有广泛的应用前景,可以提升工作人员效率,降低交通运营成本。该系统将虚拟现实技术应用于轨道交通,通过基于多视角视频的目标识别与虚拟化身生成技术构建地铁驾驶舱虚拟孪生系统,实现全方位监控和感

知能力的提升。通过地铁驾驶舱虚拟孪生系统,用户能够对虚拟驾驶舱进行全视角浏览,实现了对地铁驾驶室的全方位监控,进一步完善了整个地铁的数字孪生系统,为以后构建集感知、调度、管理、应急及服务于一体的地铁数字孪生平台奠定了基础。EA