

元宇宙（3）：软件组件

临菲信息技术港

认知幻觉在沉浸于物理空间的客观现实和用户感受到的主观现实中起着至关重要的作用。有两种类型的认知：静态认知和动态认知。静态认知是本体感觉（如视觉、听觉和触觉），而动态认知是感觉平衡和身体运动。在动态认知中，适应、注意和行为是重要的特征。

根据认知的对象，可以分为对环境的认知和对某一对象的认知。特别是在元宇宙中，减少检测和识别的失真很重要。减轻失真的方法包括改变内核的形状、改变表情和增加输入。对象识别的对象包括人脸、姿势、手势和与身体。上述这样的对象识别要经过感应、记录、识别和跟踪等过程。

场景和对象识别

对象识别是根据距离来识别物体的大小、形状、位置、亮度和颜色的过程。对于场景识别和对象识别，如上图中所示，使用了新颖的方法（如模态对齐、跨模态关注、点云和场景图）。场景识别也是对当前场景是什么状态以及它有哪些组件和配置的良好识别。在基于子图的场景图生成中，采用了一种通过聚类和共享表征将物体对聚类成图的方法。场景图是一种很好的方法，可以补充作为神经网络模型的局限性而出现的可解释属性。一些研究利用生成方法和场景图来对重叠情况下的身体进行分类，并预测人在墙后的姿势。

物体识别与场景识别同样重要，我们必须关注以人为本的场景分析和非接触性的交互（例如，凝视、手势、姿势）。当使用单个物体检测来识别许多物体时，计算的数量与物体的数量成正比，因此人们试图通过使用抽象的概念来减少计算的负担。特别是，一些研究（如 world models 和 MONET）将多个物体抽象为代表，以实现快速的物体识别和有效的训练。

声音和语音识别

识别声音和处理语音有助于了解元宇宙中的环境并方便与其他化身进行交流。而对话则是与其他化身交流和向元宇宙中的 NPC 发出指令的直接方法。由于元宇宙的连接是在不同的环境中进行的，因此拥有一种技术来分离周围的噪音和自己的声音同时还不产生噪音是非常有必要的。此外，根据距离的不同，声音的响度也成为了一个变量。进而在元宇宙中创造一个真实的环境，需要考虑到周围环境的语音识别技术，同时还要根据距离来调整音量。

场景和对象的生成

在元宇宙中生成环境和物体的方法分为反映现实世界的描绘方法和创造新的想象环境的方法。

反映现实环境的一个很现实的方法是在真实世界中再现现实世界中著名的地方（如博物馆、埃菲尔铁塔）和个人熟悉的地方（例如，家，学校）。或者，创造一个难以到达的环境（如水下、火星），来提供超现实的体验。在这样的过程中，人和物就是对象生成的主要对象。对象生成模块就负责创建一个环境中对话时任何所需人类形状的化身和 NPC（例如，名人、家庭成员）。这个模块专注于化身的面部表情和关节的自然运动，实现更流畅的多模态对话。同时，对象生成模块可生成逼真的物体，逼真到物体的细节足以让人感受到现实中存在的物体的纹理。另一种类型的对象生成则是针对想象中的动物（如独角兽、龙）和不存在的拟态对象（如会说话的椅子）。

声音和语音合成

声音合成是一个能给用户带来沉浸感的领域，但与视觉相比，就目前而言研究量还不够。它在空间中可以创造了声音，给人一种在现场的感觉，增加了沉浸感。在元宇宙中的每个角色的声音都是表达角色性格的一个重要手段。目前使用最多的是 Tacotron，这是一个基于深度学习的端到端语音合成模型。它的重点是让用户可以使用 Prosody 来强调单词或表达不确定性。而 Prosody 是一种韵律，是语音信号中与音素、说话人以及通道影响无关的变量，它可以捕捉到有意义的词语，并通过消减方法来转化他们的意思。

运动呈现

卷积神经网络和全局上下文编码被用来捕捉实时多方三维运动和姿势估计时对象之间的不对称依赖关系和上下文模式。这种方式便于反映人体的结构特征，以便在人体叠加时更准确地解释动作含义。当然也可以用单色相机捕捉困难场景的实时三维运动，并分离出人体结构（如握手），但在捕捉亲密互动（如拥抱）方面目前仍有局限。

本部分软件组件介绍已完结，请关注下一期《元宇宙（4）：内容》



临菲信息技术港



临菲信息技术港（公众号）



腾讯 • 临菲课堂



临菲编程（公众号）