

硕 士 研 究 生 读 书 报 告



题目 用于人物和场景交互的神经状态机

作者姓名 陈浩华

作者学号 22251010

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 二○二二年十二月十五日

Neural State Machine for Character-Scene Interactions[1]

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Li Qilei

By

Chen Haohua

Zhejiang University, P.R. China

2022.12.15

摘要

论文重点探讨了在达成高级场景理解的基础上，如何准确适应集合关系多样变化的交互对象。对于监督学习而言，即使是例如坐在椅子上这样一个简单的任务也是相当困难的，因为这样的任务往往涉及复杂的规划，包括周期性或非周期性运动中如何对场景几何作出反应，并精确定位、定向。基于此，论文中提出了一个名为神经状态机的深度自回归模型，它可以完美地从数据中建立多模态场景交互行为；并且还结合了一种数据增强方案，以随机切换3D几何图形，从而让角色能适应各种各样的几何场景；此外，还引入了一种结合自我中心推理和目标中心推理的双控制系统，用于提高实现目标的精度。

**关键词**：神经状态机；双向控制系统；体积表示；数据增强；动作捕捉

Abstract

The paper focuses on how to accurately adapt to interactive objects with diverse and changing set relationships on the basis of achieving advanced scene understanding. Even a simple task such as sitting on a chair is quite difficult for supervised learning because such tasks often involve complex planning, including how to respond to scene geometry during periodic or non-periodic motion, and precise positioning and orientation. Based on this, the paper proposes a deep auto-regressive model named *Neural State Machine*, which can perfectly establish multi-modal scene interaction behavior from data. And also incorporates a data augmentation scheme to randomly switch 3D geometry , so that the character can adapt to a variety of geometric scenes. In addition, a bidirectional control system combining ego-centric inference and goal-centric inference is introduced to improve the accuracy of achieving the goal.

**Keywords：**Neural state machine; Bidirectional control system; Volumetric representation; Data augmentation; Motion capture

1引言

一个角色和它所在环境的关系很大程度上决定了角色动作的真实度。以“坐在一个椅子上”这样的行为为例，角色首先要找到能绕过诸如桌子等障碍物的路径，其次也要能精确定位它相对于椅子的精确位置和方向。至于其他类似“打开门”、“搬运物体”等涉及复杂动作的任务，都面临着和上述例子相同的挑战——基于对场景的理解进行精准控制。

使用深度监督学习的高质量运动生成已经在有限的运动领域取得了成功[2][3]，但将其扩展到日常的交互场景中仍是一个难题，因为这需要模型能精准的适应更加广泛的3D几何。

论文提出的神经状态机模拟了广泛的周期性和非周期性运动行为，这一神经结构通过学习不同任务的不同运动和转换，从而获得一个用于角色控制的状态机。在此基础上，只需用相位的运动来设计控制信号，并编码成一系列高级动作的标签和目标位置，神经网络就会通过这些高级指令，学习生成精确的高质量动画。

相较于以往强制使用一个固定的相位函数去因子化网络的权重值，神经状态机是以一种端到端的方式，从运动捕捉到的数据中进行学习并因子化网络权重，从而避免了学习结果仅适用于循环的运动。

此外，为了克服训练过程中累积的误差，神经状态机使用了以自角色为中心和目标为中心的双向控制，并将两者结果结合再反馈给网络以获得更平滑更精确的动作。

另一方面，为了克服以往基于高度图的表示方法，论文引入了体积表示法来理解环境，这样就能让系统学习适应凹形物体的运动，例如穿过墙壁上的洞或从侧面滑到椅子上以避开桌子；同时还使用了数据增强，在每一帧都保持运动和交互内容的条件下，随机几何几何体，从而使得在相同数据大小下能有着更好的泛化性。

**2 神经状态机**

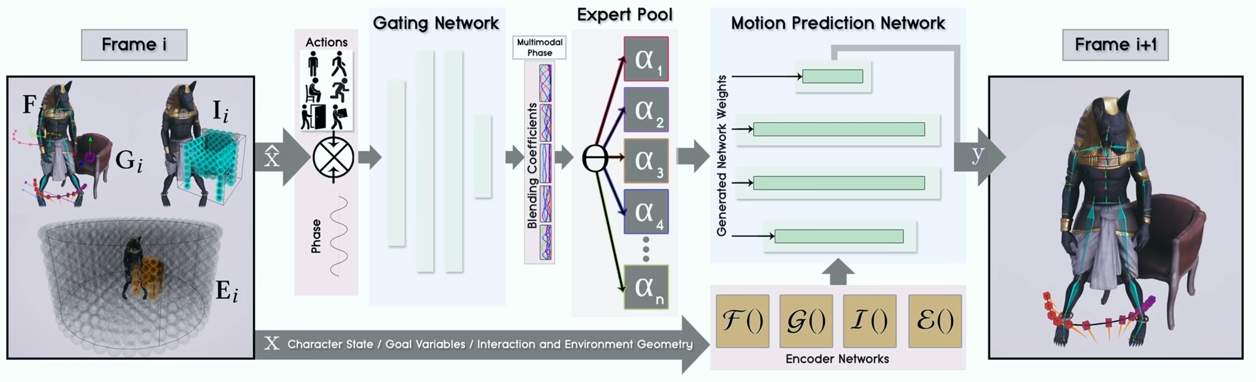


图2.1 运动预测网络和门控网络

**2.1 运动预测网络**

运动预测网络是神经状态机的主要部分，它根据前一帧的姿势，环境的几何情况，以及包括目标地点和动作状态的高级指令，来计算当前帧的角色动作。

动作预测网络由两个模块组成。模块一是个接收前一帧角色状态并将它们编码输入给一个简单三层网络的编码模块；模块二是个接收编码模块的输出结果并预测角色当前帧状态的预测模块。预测模块也是三层结构的网络，它的权重值随着门控网络的系数和一组专家权重的结合而动态更新。

运动预测网络第i帧的输入Xi表示如下：

Xi = {Fi , Gi , Ii , Ei }

Fi 是前一帧信息的输入，它是个6维的量，主要包括角色的姿态（23个关节的的位置、旋转值，与相对根坐标的速度），前后各一秒的轨迹线数据，以及采样点上的行为标签（空闲、行走、奔跑、坐下、打开、搬运、攀爬）。

Gi是目标输入，它是个3维的量，主要包括三维的目标位置；相较于前一帧的根坐标，在二维水平面上沿着采样轨迹线的方向；以及该点的目标行为（采用独热编码表示）。

Ii和Ei则分别是交互的几何信息输入和环境几何信息编码输入，是体积表示的具体数据。

运动预测网络第i帧的输出Yi由包含以下参量：

-在角色坐标系统中预测的角色姿态（根据关节的位置、旋转与相对于根坐标的速度）

-在角色坐标系统中的未来的（1秒内）根的轨迹

-在目标坐标系统中的未来的（1秒内）根的轨迹

-角色坐标系统中的关节位置预测结果

-目标数据的更新值  （更新后的位置、朝向与目标行为）

-关键关节的接触点标签（关键关节即臀、双脚与双手）

-当前的相位更新

上述输出的结果会被反馈给输入，或和用户的输入结合并作为下一轮迭代的控制信号，例如当前帧的角色姿态预测或用于下一轮迭代预测下一帧的姿态。

**2.2 门控网络**

门控网络是个三层的全连接网络，它的输出代表专家权重的混合系数。

在第i帧门控网络的输入如下表示，其中 ⊗ 为Kronecker积运算：

Xi = Pi ⊗Xi′

第一个向量Pi是一个二维的相位向量，它是该帧上所标记的相位标量：

Pi = {sin(pi ), cos(pi )} ∈ R2

Xi′则包括了当前的行为标签、目标的空间位置、目标朝向、目标行为，同时研究发现，取目标之间的距离标量δ ，以及与目标朝向之间的角标量θ ，将二者各自与目标行为作乘，可以在多任务的运动合成中获得较好的表现，对于靠近目标位置时的足步规划尤甚。

上文提到，相较于以往的固定相位函数对网络权重进行因子化，门控网络结合混合系数动态调整相位函数，直接表现就是滑步、运动突变的效果更真实。

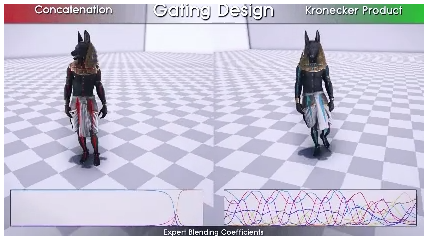
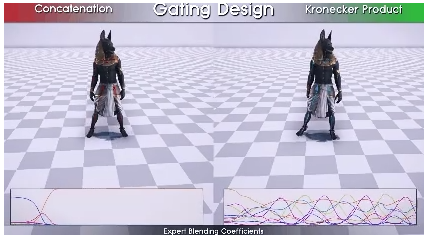


图2.2 固定相位与动态相位函数的比较

**3 目标驱动运动合成**

**3.1目标驱动的角色控制**

在本研究中，由两种角色控制方式，分别是高级的目标驱动方式和低级的移动驱动方式。在目标驱动中，用户只需要用鼠标点击目标物体（椅子、桌子）或者键盘输入想要执行的动作，角色就会自动完成指定任务；在运动驱动中则是和常规的键盘控制行走、跑步类似。



图3.1 移动驱动（左）和目标驱动（右）

这两种控制模式之间可以通过目标输入Gi来实现无缝切换：在移动模式下，将目标输出Go与未来1s内用户输入的控制信号混合，并将混合结果作为 Gi 再次输入网络；在目标驱动模式中，则以用户所选的物体或者环境作为 Gi。它可以随时改变，并基于角色姿态和行走模式来生成动作间的过渡。

例如图3.2所示，系统将“坐下”的指令输入门控网络，网络计算需要混合的专家权重来生成下一帧的运动，随后运动预测网络根据这一结果来逐帧生成行为标签。

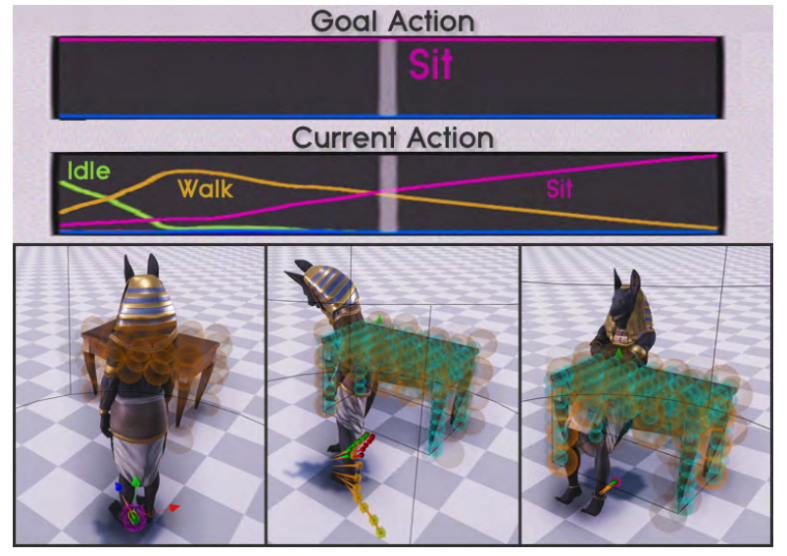


图3.2 “坐下”指令

**3.2双向控制方案**

双向控制器的核心思想在于同时从主体中心点和目标中心点来预测运动，并将预测结果反馈给神经状态机，来增加角色靠近目标的精确度。当仅使用主体中心控制时（这也是当下常用的实时角色控制器），很难到达目标并执行任务。



图3.3 双向轨迹预测：红色轨迹为Root空间，绿色为目标空间

这是因为基于主体中心的控制依赖于Root的实时更新，而多帧运算后会存在误差积累，故不能保证角色准确停在目标位置。反之，在目标空间中获取运动的预测信息能够获取更精确的值，将其反向转换到Root空间，则可以避免误差的积累。尤其在训练数据量不足的情况下，从随机位置运动到目标点，使用双向控制方案能够获得更高的精确度。

**3.3体积感应器**

体积感应器由环境感应器和交互感应器组成

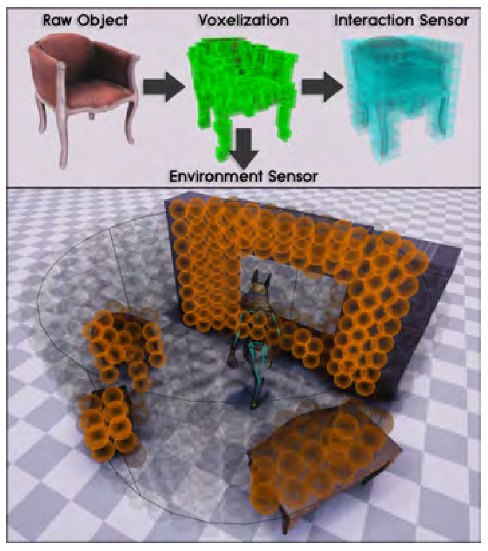


图3.4 交互感应器（上）环境感应器（下）

**3.3.1环境感应器**

环境感应器是个半径为R高为H，并以角色为中心的圆柱体。当角色移动时，会评估圆柱体与物体或者环境的碰撞关系 ，并记录在环境几何信息 Ei 中。在圆柱体内，对半径为 r << R 的球体进行采样，并测试他们与物体或者环境的相交部分。球体采样采用极坐标样式，沿着半径、弧度和高度方向的采样步长分别为([R/2r], [πR/r], [H/r])。

对于每个单位球体，大致计算其被物体或者环境中的碰撞体所占据的体积，并以浮点形式输出。若单位球体中心在碰撞体中，则输出 1 ；若与碰撞体无交集，则输出 0 ；否则输出 1−d/r 。其中 d 为碰撞体距离单位球体的最近距离。这种设计能够为体积感应器提供连续的输入，进而产生平滑的动作。所用的碰撞体源自高分辨率体素模型的自动计算，而感应器的值会输入进 Ei 。

**3.3.2交互感应器**

交互感应器是为了提高交互的精度，例如环境感应器可以感应到障碍物所在，但无法具体到诸如角色的手臂如何放在椅子上。

交互感应器用8×8×8的立方体集群覆盖整个目标物体，每个立方单位包含一个四维向量，向量中前三个值表示立方体中心和物体根部的相对矢量，最后一个数值同环境感应器中计算碰撞体所占据的体积。这些数据最后被提交给交互几何输入Ii。交互感应器可提高动作的精度，从而实现例如搬运箱子等的复杂动作，具体可根据目标输入Gi来动态调整交互感应器的位置数据。

**4数据准备与实验**

数据准备阶段分为三个部分：动捕、标记和数据增强。

研究中使用了XSens惯性动捕系统，输出了共计94分钟的动捕数据，并通过镜像操作使数据量翻倍。动捕的内容包括了靠近一把椅子并坐下、行走避障、打开门并通过、拾取物品、携带箱体再放下。除了这些基本动作，研究还涉及了一些复杂的场景。比如针对有一部分椅面在桌下的情况，角色需要弯曲双腿，用手支撑着桌面来完成整个坐下的动作。

每帧的数都会用一个行为标签记录。行为标签包括空闲、行走、奔跑、坐下、打开、搬运、攀爬等单一标签，也存在“搬运+空闲”或者“搬运+行走”这类复合标签。除了标记当前帧正在进行的行为，还会为当前帧标记一个目标行为，即通过目标输入Gi来让训练系统结合当前行为和未来行为一起训练，从而可以让用户能和控制角色实时交互。对于行走等周期运动，会根据左-右-左脚面着地规律，将其相位定义为0-π-2π，中间部分则采用插值进行过渡。对于非周期运动，0与2π分别为动作过渡的开始及结束帧，中间部分的相位标量仍进行插值处理。

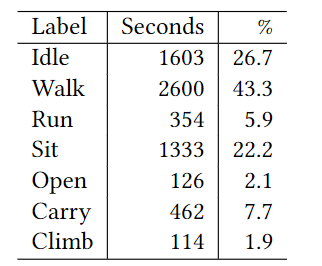


图4.1 不同行为标签的测试数据

为了在不增加动捕数据的基础上提高结果的普适性、覆盖到更多不同形状的交互物体，研究中设计了一个可以保留交互上下文的数据增强框架。先用一个基础模型去匹配该段动捕数据，再逐帧随机地从池中采样新模型（池中数据源自ShapeNet[4]）,随后根据Relationship Descriptor的思路[5]去匹配角色姿态。

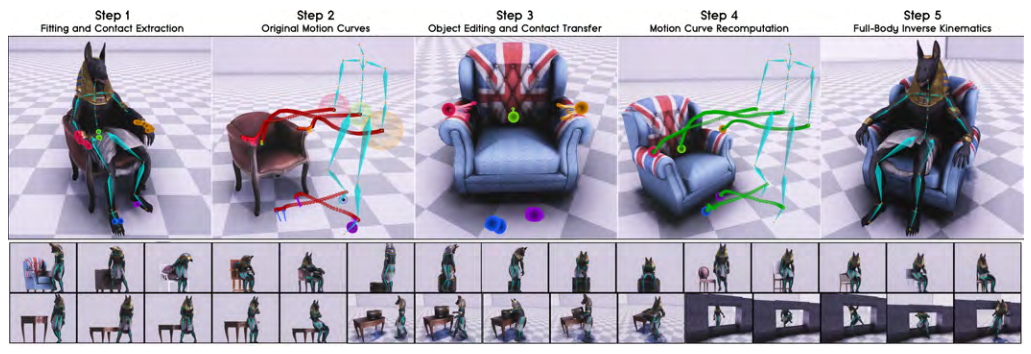


图4.2 数据增强的五个步骤

具体步骤如图4.2所示。(1)选定该段动捕数据中角色与物体相对静止的帧，指定好基础模型的位置和朝向，使其匹配当前角色姿态；(2)当关键关节（如图中的双手、双脚与臀部）与物体或者环境的距离小于一定阈值时，物体或环境的表面会生成接触点。关键关节的位置会被重新定义为从对应接触点发出的矢量的末端；(3)切换模型，或者对模型进行空间变换。接触点坐标会通过跟随变换、近表面投影、手动指定等方式重置；(4)根据新的接触点，更新关键关节的轨迹；(5)使用基于CCD的全身IK，重新计算角色姿态。为了减少终端感受器的误差，从根部开始逐级向下计算位置与朝向。这种方法无法保证时间连续性，但能得到稳定且平滑的结果。

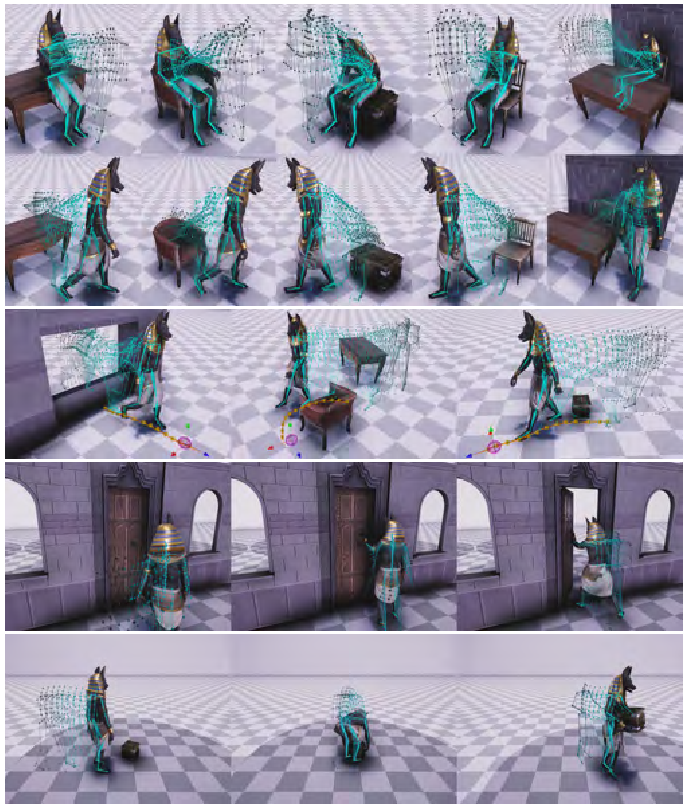


图4.3 多种行为的训练过程

**5 小结**

基于神经状态机的交互动画效果很精美，但也存在一定缺陷：

1. 研究中使用的交互几何体数据较小，对于差异于训练数据集中的物体时，角色难以与之进行交互；也许使用其他几何表示法或者PointNet++[6]这种用ShapeNet预训练过的几何编码器可克服一定问题。
2. 切换目标行为会产生动作突变。由于网络输入变化不连续，改变目标行为时，容易造成角色的运动突变。
3. 角色的交互类型有限，目前仅限于坐下、托举、避障等交互类型
4. 非周期运动的相位标记需要人工成本

但从论文中实验的成果来看，它提出了一种新的涉及与环境交互的神经网络框架，并允许用户提供抽象目标来引导角色，且可用于实时应用，依旧是很大的突破，或许随着数据集不断扩大，将来就能真正应用于其他的场景中，这是值得让人期待的。

参考文献

1. Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. 2019. Neural state machine  
   for character-scene interactions. ACM Trans on Graph 38, 6 (2019), 209.
2. Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned neural networks  
   for character control. ACM Trans on Graph 36, 4 (2017), 42.
3. Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for  
   character motion synthesis and editing. ACM Trans on Graph 35, 4 (2016)
4. Chang, Angel X., et al. "Shapenet: An information-rich 3d model repository."arXiv preprint arXiv:1512.03012(2015).
5. Al-Asqhar, Rami Ali, Taku Komura, and Myung Geol Choi. "Relationship descriptors for interactive motion adaptation."Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. ACM, 2013.
6. Qi, Charles Ruizhongtai, et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space."Advances in neural information processing systems. 2017.