

硕 士 研 究 生 读 书 报 告



题目 基于以活动为中心的功能性三维场景建模

作者姓名 徐 克

作者学号 21551072

指导教师 李启雷

学科专业 移动互联网与游戏开发技术

所在学院 软件学院

提交日期 二○ 一五 年 十二 月

Activity-centric Scene Synthesis for Functional 3D Scene Modeling

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

XuKe

Zhejiang University, P.R. China

2009

摘要

我们提出了一种新的方法来生成3D场景，它允许在存有噪音和不完整的3D扫描捕捉的情况下，生成和正式场景一样的活动。从低质量的深度数据中进行可信的目标检测和实例检索是具有挑战性的，我们的算法旨在建立正确的语义模型，而不是精确的排列几何对象。我们的核心贡献是使用一个新的场景合成技术，在再现一个粗糙的几何场景的情况下，从场景数据库中匹配并使用功能类似的场景模型。基本上我们的场景合成方法的主要观点是，许多现实世界环境中的结构设计可以方便特定的人类活动，如睡觉或吃饭。我们使用它们经常被使用的一些活动通过虚拟代理和对象排列之间的关联代表场景功能。当对一个场景进行建模时，我们首先通过一个被扫描的环境来确定该环境所支持的活动。然后，我们对从被观察到的场景几何形状的约束数据库中检索虚拟对象，并对其进行合理的语义安排。对于给定的3D扫描，我们的算法产生一系列的合成场景，它们支持所捕捉到的真实环境中的活动。在感知评估研究中，我们证明了我们的结果被判断为在视觉上的吸引力和功能性方面媲美手动设计的场景。

**关键词**：场景合成，活动，场景理解

Abstract

We present a novel method to generate 3D scenes that allow the same activities as real environments captured through noisy and incomplete 3D scans. As robust object detection and instance retrieval from low-quality depth data is challenging, our algorithm aims to model semantically correct rather than geometrically accurate object arrangements. Our core contribution is a new scene synthesis technique which, conditioned on a coarse geometric scene representation, models functionally similar scenes using prior knowledge learned from a scene database. The key insight underlying our scene synthesis approach is that many real-world environments are structured to facilitate specific human activities, such as sleeping or eating. We represent scene functionalities through virtual agents that associate object arrangements with the activities for which they are typically used. When modeling a scene, we first identify the activities supported by a scanned environment. We then determine semantically plausible arrangements of virtual objects retrieved from a shape database constrained by the observed scene geometry. For a given 3D scan, our algorithm produces a variety of synthesized scenes which support the activities of the captured real environments. In a perceptual evaluation study, we demonstrate that our results are judged to be visually appealing and functionally comparable to manually designed scenes.

**Keywords：**scene synthesis, activities, scene understanding

1引言

对于大多数应用来说，对现实环境中的物体进行标识、排列布局和理解其功能是一项困难但很重要的工作。这项工作之所以困难，是因为真实的环境只能由观测到的不完整的视点集合捕捉，并且每个观测点都可能受到遮挡和传感器质量的限制。尽管如此，人们依然能够通过映射他们以前经历过的环境，清楚地了解周围的环境。

与人类相反，计算机很难理解物体和他们的3D场景布局。一个显著问题是在识别物体时，当对象的一部分是可见的，或当一个确切的几何匹配并不在训练数据库中时，工作将变得很困难。一个有前途的研究方向用更高级别的上下文信息，来增加低级别的几何特征，从而克服这些限制。不幸的是，在实际设置中对语义语境建模是具有挑战性的，因为它很难知道哪些对象之间的上下文关系是重要的。我们希望通过以活动为中心使环境中的功能解析，我们的观点是用人类的活动作为一个隐藏的变量来表示对象之间的上下文关系。它能够使一个简单而稀疏的情景再现。我们的做法是由于真实场景通常是考虑到特定功能被设计，例如卧室用于人类睡眠，而在办公环境中能够方便地使用书桌和电脑。虽然我们评估我们在3D场景中生成任务的方法，但我们相信理解活动对于许多无法进行场景重构的领域是至关重要的，包括图像理解和人机交互。

在这篇论文中，我们使用在一系列的活动中观察到的场景，以相关的活动区域为中心引导场景建模。我们采用环境中的RGB-D相机进行扫描得到的数据生成一组3D模型作为输入，其允许相同的交互类型作为扫描输入。这些活动区域引导场景建模朝向一个合理的结果，更重要的是，它们允许遮挡和噪声数据的存在。

**2 相关工作**

因为我们的目标是以人类为中心来重建场景，因此，了解人类活动如何影响场景的构建是非常重要的，它与认知心理学有着密切的联系。

通过语义重建三维扫描场景的一种方法是从几何形状数据库中检索模型，一旦对象可以被检索到，便可以从数据库抽取更高级别的语义信息，并用此来解释三维场景。不幸的是，基于低层次几何特征的物体识别和检索在数据存在噪音和范围未分段的情况下是具有挑战性的，最近的方法是结合局部与整体的特征描述来共同对物体进行分段、检索和排列。另一种方法是通过机器学习识别模型，根据物体的几何特征，对其进行分类以检测物体。从多个范围的图像中得到模型的变形部分，使用随机决策森林通过对高质量范围的数据过度分割对对象进行分类。根据环境中的立方体集合，可以通过立方体的排列布局来分析场景结构。

为了提高重建的效率，前人增加了对象实例的检索与空间上下文信息，如果传感器数据是不完整的，或者是存在噪音的，这将是非常有意义的。我们的主要目标是对语义合理的场景建模，我们的方法是对扫描输入的几何体使用较少的约束，不集中于识别对象实例，而是放宽精确的实例匹配要求，来捕捉对象间的布局和关系。

上下文场景理解对于创建自动化内容来说，已经变的同用于组织和维护的场景数据库一样重要。一般使用关系曲线图对同一场景中的不同物体构建上下文关系，Xu等人在2014年将集群场景关系图转化为上下文重点组，使他们能够简化图形连通性，从而更有效地组织大型场景的集合，在大型场景集合中，Liu等人 在2014利用概率性事件来获得一致的图形层次结构。我们的方法扩展了这些想法，具体而言，我们通过隐藏的代理变量间接地表示在活动中存在的关系和模型对象之间的关系，这使我们能够最小化数量关系，因为我们只能经由代理考虑高层语义含义，并避免必须通过识别对象实例来推断场景的关系。

[Matthew Fisher](http://graphics.stanford.edu/~mdfisher/publications.html)等人在2012年从场景数据库和用户提供的输入样本中推导出贝叶斯网络，并通过贝叶斯网络对这些对象关系建模，生成合成场景。Xu等人在2013年提出了另一种场景合成系统，他们使用简单的图像草图来驱动合成算法。虽然我们也使用场景数据库进行合成，但我们在建立场景模型之前，使用现实世界的RGB-D扫描的多个活动作为前提。

**3 准备工作**

给定一个三维扫描，我们首先需要为捕获的场景构建相应的场景模板。一个场景模板是一个场景同时用其几何和活动特性进行编码的高级表示，它是输入扫描对象和输出场景对象的中间表示。我们的场景模板被定义在一个已知范围的固定矩形空间中。场景模板需要捕捉环境中几何布局的大致表示，重点是引导合成结果朝场景布局逼真的方向发展，并且需要捕捉场景中对象的属性，例如某个家具是否存在或物体上的表面的密度和高度，但不是特定属性，例如一个显示器的品牌或一本书的标题。

构建场景模板的具体做法是：

1.首先确定场景几何描述

将一个场景划分为垂直排列的均匀栅格，各覆盖5厘米×5厘米。在每个列c，我们存储两个距离：CS：地面上最高支撑平面的高度，CR：最高支撑平面上方观测到的最高几何体的高度。支撑平面指支撑其他物体的平面，如桌子的表面，床的表面，或咖啡桌的表面。划分支撑面和位于支撑面上的几何体对于捕捉对象的大致位置是有用的，我们使用的TG来指代（CS，CR）高度值所需的网格为场景模板T。如图1所示，蓝色列表示支承面在该区域的期望高度（CS），绿色列表示坐落在支承面上面的剩余几何体的高度（CR）。从一个场景模板中生成场景将尝试从区域中的所有垂直列匹配支撑面和剩余几何体的高度。

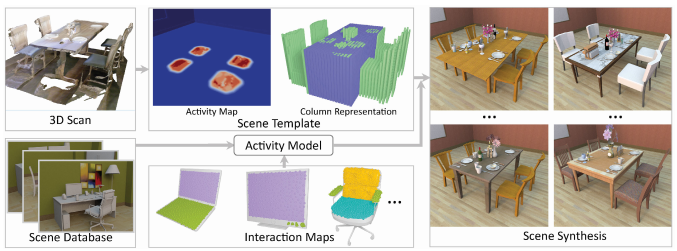


图1. 场景的几何描述

2.确定场景的活动描述

我们在一个场景模板中表示的多个活动作为一组活动映射，它们是连续分布定义在二维平面上的一组活动。我们使用表1给出的这组活动，一个场景模板定义一组密集的活动样本（X，Y，θ），范围从0％（委托代理几乎不可能执行的活动）到100的％（委托代理极有可能执行的活动）。（X，Y，θ）的三个分量分别代表执行活动可能的位置和方向。这种表示方法从场景模板生成场景时将尝试定位对象，从而执行在活动映射中指定的活动。

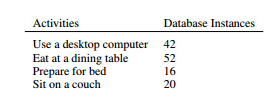


表1

1. 取样委托代理

当进行活动映射，我们开始通过采样持续性的活动离散分布的委托代理集合的TA。（X，Y，Z，θ）坐标表示代理人的头部预期位置和方向，通过活动标签和代理人的头部位置可以识别一个委托代理，代理人的头部位置隐性制约了身体其他部位的期望位置，如肩膀为相对于代理的头部和面部的方向固定偏移量。对委托代理进行取样，首先在前10％的样本中选择一个随机样本作为这个位置的委托代理，委托代理的高度被作为在场景数据库中进行活动观察的平均高度。图2表示了从一组活动映射中对代理取样的一个例子。

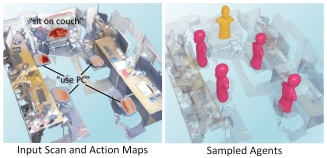
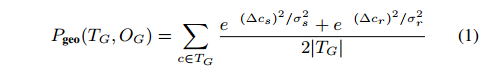


图2. 取样委托代理

接下来介绍如何创建几何模型和活动模型。建立几何模型时，为了量化合成的对象布置O与的一致性，我们通过计算—类似于合成场景中列的网格，从一个分类的三维模型数据库中得到O中每个对象。通过查看现有场景数据库，确定将某些类别作为分类支持，并查找已观察到支持至少两个其他对象的所有类别。为了计算，我们从顶部各列向下拍摄场景。

为了比较和，对所有列c的比较函数取平均值：



ΔCS和ΔCR是支撑平面之间的高度差和剩余几何体之间的高度差。σs和σr表示几何体合成的程度。 Pgeo（，）范围是从0（几何相似度为0）到1（精确的几何匹配）。

我们想要评估一个给定的目标配置能对一个场景模板中的每个活动映射提供怎样的支持，还需要确定对活动映射采样生成一组代理后，每个代理是如何执行自身所标识的活动。为了实现这一点，我们使用了一个语义数据库来为每个活动建立一个模型。使用这些样例来表示对象的预期分布和相对于代理对象的预期方向和位置。我们的活动模型还包含了对活动进行展示时代理与每个对象交互方式的相关内容。例如，一些对象可能含有代理想要看到或触摸的区域。

**4 场景合成**

我们的目标是根据给出的场景模板合成具有真实性的场景，我们的最终任务是生成大量相似的候选场景，然后使用一个简单的贪心算法从候选的场景集合中选择估算结果最高的场景。场景模板的估算函数需要输入一个场景模板T和对象布置O，输出结果为0（完全不同）到1（完全一致）之间的值：

|  |  |
| --- | --- |
|  | （2） |

其中，声明对象所需的活动位置和方向，其中该代理将能够有效地与它们进行交互。，只有在模板的每个代理均有可绑定的一组对象约束时返回一个高分评估，在评估函数中，它不需要被明确建模，因为我们将直接从每个活动的对象分布进行采样。这确保了在模板的每个代理均有可绑定的一组对象约束。为了生成候选场景，我们首先取样了一组委托代理使用前面描述的方法做出现场模板的连续动作分布。反复生成候选场景，通过使用一个简单的贪心算法进行场景模板匹配来生成场景。我们根据公式（2）生成一个候选的场景，根据估算函数计算匹配分数，选择得分最高的。

接下来，我们使用每个活动贝叶斯网络将一组对象的样本接入每个代理。然后，我们开始反复建立候选模型列表和位置密集采样列表并将其插入，使用贪心算法选择最佳候选模板生成场景。合成算法的伪代码在图3给出。

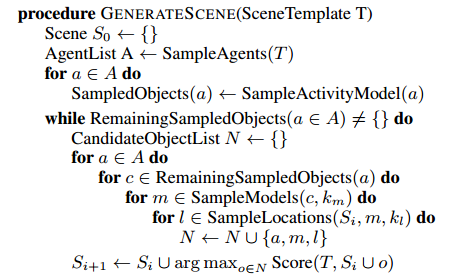


图3. 合成算法的伪代码

为了评估我们的工作，我们使用了微软的Kinect传感器来扫描一组杂乱的室内场景。它支持的活动在表1中列出。我们使用的体积融合算法来对齐扫描的RGB-D-帧和重建三维网格，将此作为我们方法的输入，如图4左侧所示。图4显示了两个扫描的场景以及用我们的方法合成的场景。



图4.左图：输入3D扫描；右：用我们的方法在不同的环境中建模的结果

**5 总结**

我们已经探讨了如何将一个在场景模板中编码的活动作为焦点来表示三维场景的语义的方法。我们相信对活动的深刻理解对于我们理解场景和任务重建是至关重要的。复杂的活动，如“油炸鸡”，涵盖了广泛的子活动，并与潜在的数百个对象共享广阔的空间。厨房的结构使得这些活动可以很容易和有效地执行。这符合我们为了能存在于其中而创造的精确环境。我们所探讨的这些活动，与他们真是生活的部分相比，已经得到了很大程度的简化。现实世界中的代理以及定位和对象的相互作用，可以模拟复杂的多级动态如打开橱柜或重新布置家具。

我们也评估了一个代理如何以其自身的能力来展示一个活动，以便以特定的方式同与之相邻的对象交互。我们的活动模型来源于虚拟场景里的一个扩展数据库，其中每个对象都是手动标记的区域，这些区域允许不同类型的相互作用，从这些虚拟数据库中学习的挑战是，他们是静态的--在这些环境中执行的动作、代理、动态运动必须通过手动的近似化和注解。尽管如此，我认为我们的工作为那些支持更一般化的活动探索提供了一个坚实的基础。

参考文献

[1][Matthew Fisher](http://dl.acm.org/author_page.cfm?id=81332498586" \t "http://tog.acm.org/_blank), [Manolis Savva](http://dl.acm.org/author_page.cfm?id=81487650815" \t "http://tog.acm.org/_blank), [Yangyan Li](http://dl.acm.org/author_page.cfm?id=90658669257" \t "http://tog.acm.org/_blank),[Pat Hanrahan](http://dl.acm.org/author_page.cfm?id=81100482576" \t "http://tog.acm.org/_blank), [Matthias Nießner](http://dl.acm.org/author_page.cfm?id=81464672669" \t "http://tog.acm.org/_blank). Activity-centric Scene Synthesis for Functional 3D Scene Modeling． New York, USA.Volume 34 Issue 6, November 2015  Article No. 179 .

[2]Manolis Savva,Angel X.Chang,Pat Hanrahan, Matthew Fisher, Matthias Nießner. SceneGrok: Inferring Action Maps in 3D Environments. SIGGRAPH Asia 2014, Shenzhen.