基于 Kinect 的深度数据融合方法*

吴 维，胡 涛，朱欣焰
（武汉大学 测绘遥感信息工程国家重点实验室，武汉 430079）

摘 要：与传统三维激光扫描仪相比，Kinect 作为一种新型深度相机，具有价格低廉、深度数据获取能力强、RGB 影像与深度影像同步获取等优势，然而面对较大室内场景精细建模却存在数据量大、建模范围有限、对硬件环境依赖性强等问题。因此，在现有单一模型建模基础上，提出了基于 Kinect 深度影像的多模型数据融合方法，实现模型间的自动拼接。最后通过两组实验对提出的数据融合方法进行了验证，并取得了较好的模型拼接效果。

关键词：Kinect 传感器；深度影像；体元数据；数据融合；模型拼接
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2014)01-0285-04
doi:10.3969/j.issn.1001-3695.2014.01.067

Kinect-based depth image fusion method

GUO Wei, HU Tao, ZHU Xin-yan
(*State Key Laboratory for Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University, Wuhan* 430079, *China*)

Abstract: Compared with traditional 3D scanners, Kinect is a new kind of range camera, it has advantages of low price, strong capability of data collection, synchronous acquisition of RGB image and depth image. However, it exists problems of large amount of data, limited reconstruction range, and high dependence on hardware, when Kinect is used to build large and fine indoor scenario model. Therefore, this paper proposed an effective solution to reconstruct high-resolution 3D model in large indoor space and it was especially to solve the problem of data fusion with different Kinect image data in the same scene and merging with different models automatically. Finally, two sets of experiments verified the proposed method and the result proved that it was very effective.

Key words: Kinect sensor; depth image; volume; data fusion; model merging

0 引言

随着 2010 年微软公司开发的体感设备 Kinect 面世，相比 LiDAR、高清相机等传统常用室内测图设备，Kinect 具有以下优势：a) Kinect 具有廉价轻便的特点，适合室内拥挤场景和自发地理信息采集模式；b) 同步获取场景深度与影像信息，可全自动完成场景纹理映射；c) Kinect 以 30 fps 速率连续输出，可获取大量深度信息，便于保持采集过程中信息的完整性。

目前，Kinect 已较广泛应用于三维场景建模领域。本文将重点研究深度数据融合的方法，尤其是基于体元的数据融合，将其应用于 Kinect 深度影像，重建三维模型，并分析其中所存在的不足。然而，仅依赖于体元重建精细的三维模型会造成十分严重的内存资源消耗，无法重建出大型的三维场景，因此本文将扩展 Kinect 重建能力，提出多模型拼接方法，在保证重建三维模型精度的同时，重建出大型的室内三维场景。

Kinect 是微软开发的体感设备，它属于一类名为 RIM (range imaging) 相机[1]的新型传感器，主要用于捕捉人体骨骼结构，以此实现以身体作为控制器的构想。Kinect 由一个发射单元（脉冲光、调制光或结构光）、感光传感器（CCD、CMOS 或 APD）、光学系统以及一些驱动电路和计算单元组成。Kinect 设备同步获取场景深度与影像信息，可全自动完成场景纹理映射。基于 Kinect 设备获取的深度数据是结构化的点云数据，本文采用基于体元的数据融合方法重建三维模型。

体元技术最早由 Baumgart[2] 提出，并指出每一个体元只有 0 和 1 两个状态，1 意味着该体元被重建目标占据，0 则反之。1989 年 Elfes[3] 利用声纳为移动机器人导航时提出占用栅格概念（occupancy grid），它是将整个空间中的体元分为 occupied、free、unknown 三类，并利用概率函数表达空间使用情况，利用该方法可以实现机器人自主导航定位。Hoppe 等人[4] 通过构造点到物体表面的有向距离场（signed distance function）来重建物体表面。为改善 Hoppe 方法中存在的边界重建不理想的问题，并同时解决法向传播中可能出现的局部孤岛问题，Curless 等人[5] 在每个体素中保存两个值，一个是权重信息，另一个是距离值，他们将每幅深度图像转换成一个加权有向距离场（weighted signed distance function）。

在 Kinect 发布以后，华盛顿大学和 Intel 合作开展了名为 RGB-D Dense Point Cloud Mapping 的研究项目[6,7]，该项目旨在利用 Kinect 实现机器人的自动测图。由于较少关心测图本身精度，因此重建效果不够理想，场景中存在很多“洞”，且有重影现象产生。2011 年 8 月，微软在 SIGGRAPH 大会上展示了

收稿日期：2013-02-19；**修回日期**：2013-04-02 **基金项目**：国家“863”计划资助项目(2012BAH35B03,2011AA010500)；中央高校基本科研业务费专项基金资助项目(2012619020215)

作者简介：吴维(1981-)，男，荆州公安人，博士，主要研究方向为三维建模（guowei98032@gmail.com）；胡涛(1985-)，男，博士研究生，主要研究方向为模式识别；朱欣焰（1963-），男，教授，博士，主要研究方向为空间信息服务、空间数据库等.

其 KinectFusion[8,9] 项目成果，旨在利用 Kinect 实现增强现实(augmented reality)。其中针对 Kinect 原始深度图精度低的问题，利用 Kinect 高频输出特点，基于文献[10,11]提出体重构技术和 Point-To-Plane ICP 算法，采用截断符号距离函数(truncated signed distance functions)和 GPU 进行并行加速，实现实时精细三维建模，精度可达到毫米级。

虽然 KinectFusion 方法得到了很好的效果，但依然存在一些问题未解决。首先，面对大型室内场景，密集体元的表达方法需要消耗大量的显存空间。另一个重要的挑战是在较大场景测图过程中，很可能由于传感器移动过快或平面困难区而导致 ICP 算法失效，继而跟踪丢失情况且难以恢复。针对上述问题，本文提出多模型拼接方法，在保证重建三维模型精度的同时，重建出大型的室内三维场景。下面将介绍传统的数据拼接方法，并基于该拼接方法提出了针对大型场景模型的数据融合方法。通过两组实验对提出的数据融合方法进行了验证，并取得很好的效果。

1 数据拼接方法

微软的 KinectFusion 项目旨在重建出高精度的三维模型，采用的是体绘制技术，重建模模型可达到毫米级精度，而且可对模型同步映射纹理信息。RGB-D SLAM 是基于面绘制的技术，以点云的形式重构出大型室内场景，但精度不高。本文侧重于重建大型且精细的室内三维场景，因此基于体绘制技术融合 Kinect 深度数据重建三维模型，体元数据融合的一般流程如图 1 所示。

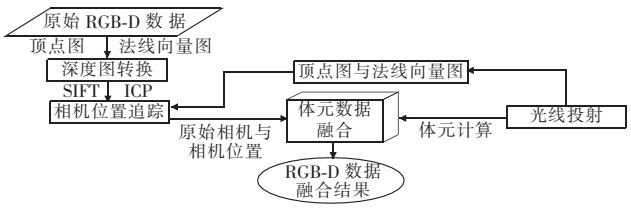


图 1 Kinect 深度数据融合过程

1)数据预处理

在使用深度相机获取深度数据之前，通常需要对这些输入数据进行预处理，这包括去噪声影响、数据转换两大块。

2)相机位置追踪

在 Kinect 设备对实体场景进行多视角扫描过程中，通过配准的方法实现相机位置的追踪，使得不同视点获取的 RGB-D 影像位于统一坐标系下，为下一步数据融合提供标准的数据结构。

3)体元数据融合

基于隐函数法将物体多个方向的表面数据合并，得到物体的完整表面，其中心思想是使用一个连续的隐式函数 $D(x)$ 来表示每一个采样点。该函数中每一个点 X 都带有权重信息和距离信息(weighted signed distance)，从体素 X 沿视线观察方向 F 与物体表面相交。通过计算有向距离函数(signed distance function，SDF)[2] 得到的距离值可正可负，正则意味着沿射线方向上该体元位于重建表面之前，负则位于重建表面之后，而正负相交处则为重建物体真正的表面。也就是说 $D(x)=0$ 就是重建出的物体表面。

随着数据不断进入，此时需要更新每个体元所记录的值，而从不同方向对同一区域采集其深度数据时，受物体表面情况和传感器性能影响，其误差不同，因此数据融合的时候应对其加权求平均。如果权重设置不当，很可能使重建出的表面出现褶皱。此时函数 $D(x)$ 的值可以通过 $\{d_i(x)\mid i=0,\cdots,n\}$ 和对应的权重信息 $\{w_i(x)\mid i=0,\cdots,n\}$ 通过式(1)计算。

$$D(x)=\frac{\sum w_i(x)d_i(x)}{\sum w_i(x)} \tag{1}$$

2 大型场景模型数据拼接方法

在第 1 章中介绍到由于 KinectFusion 系统受 GPU 内存限制，要重建出较为精细的室内三维场景，融合实时 RGB-D 数据最多只能重建出 5 m×5 m×5 m 大小的室内场景，这明显不能满足实际应用的需求。为此本文提出多模型拼接的技术，其主体思路是将每一次数据融合所得到的模型通过离线处理，将多个模型融合到一起，最终生成大场景的三维模型。其算法流程如图 2 所示。

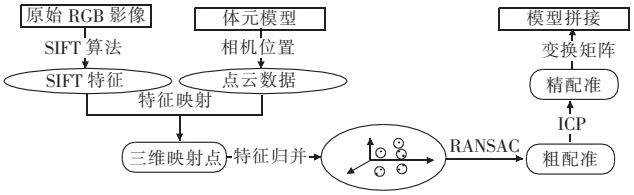


图 2 多模型拼接方法

2.1 关键帧选取

场景扫描是一个具有时间序列且不简单的过程，若选择相邻帧的 RGB 影像作为矩阵变换的输入，由于重叠区域过多，且计算量较大，将造成不必要的开销；若选择时间间隔较长的帧分别作为匹配的 RGB 影像，由于重叠区域较小，将造成匹配效果不佳。因此，如何选择合适的关键帧对于后期模型融合结果将产生重要影响。

关键帧的选择需要注意两点：一是关键帧数据具有完整性，即要求对扫描空间的每个区域都能覆盖到；二是在保证融合结果的前提下，选取尽可能少的关键帧数量。在 RGB-D SLAM[7] 中，研究人员通过计算相片之间的覆盖度来决定是否添加新的关键帧，这种方法比较消耗计算资源，而且在室内场景中，有纹理重复和稀疏的区域。因此，本文直接计算相机的角度偏转量和平移量以决定关键帧的添加，如图 3 所示。对每一帧数据，都有对应的三个旋转角度以及平移量，可以计算其相对前一关键帧数据的偏移量，当大于某一阈值时就可以将其添加为关键帧。

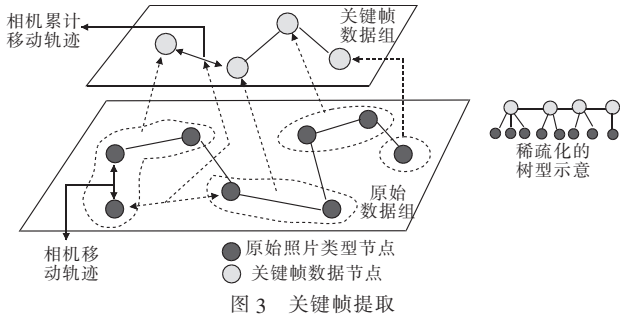


图 3 关键帧提取

2.2 光线投射

由于在上述扫描过程中只保存了 RGB 影像，Kinect 设备获取的原始深度影像质量较差，因此从体元模型中生成的深度

影像具有较高精度,此时需要利用光线投射技术。光线投射的基础是从投影中心投射出一条光线,直到该光线达到阻挡其继续传播的最近物体的表面。在此处,主要利用光线投射算法确定 RGB 影像上每个像素所对应的 *XYZ* 坐标信息,得到的坐标信息主要用于求解两个点云模型间的刚性变换。

在关键帧获取过程中,获取 RGB 影像的同时,同步获取了该关键帧对应的相机位置参数,利用该参数就可以使用光线投射算法从体元模型中获取 RGB 影像每个像素所对应的 *XYZ* 坐标。假设 RGB 影像中像素的坐标为 $\upsilon=(r,c)$,相机的内部参数是 K,相机的全局位置为 $T_{g,k}$,则该像素对应的射线为 $r=T_{g,k}K^{-1}\upsilon$,在体元中沿 r 遍历该方向上的体元,直到碰到记录 $D(x)=0$ 处的体元为止,返回此处的 *XYZ* 坐标即为该像素对应的三维坐标。

2.3　影像集数据匹配与模型拼接

对于 Kinect 重建出的每个三维模型,都对应有一组关键帧数据,本文正是利用这些数据计算得到两个三维模型之间的变换关系。这一过程主要包括特征提取、特征归并、特征匹配、计算变换矩阵、精匹配。

1)特征提取

基于现有的 RGB-D 影像,若直接利用点云数据坐标信息进行两幅影像的特征匹配比较困难,相比之下,在 RGB 影像中提取特征并匹配具有更好的准确性和鲁棒性。由于 RGB 影像与点云数据已配准,因此将 RGB 影像上提取得到的特征点映射到点云数据上就可作为点云数据的特征。基于 SIFT 算子的特征提取已广泛运用于影像匹配中,由于传统基于 CPU 的 SIFT 算子无法满足快速处理的需求,因此本文采用基于 GPU 的 SIFT 算子,可达到 11 fps 的处理速度。

将 SIFT 特征映射到点云数据上时,需要注意点云数据中存在大量无效数据,因此映射时需要对数据的有效性进行判断。一般来说无效数据出现于以下两种情况:a)深度影像的可视范围与颜色影像的可视范围不同,颜色影像上存在的区域在深度影像上很可能无法接收到深度信息;b)SIFT 特征处于物体的边缘,而点云数据在边缘处保持不好。因此这两个原因都可能导致映射到无效区域。当 SIFT 特征对应点云数据上无效点的时候,可以直接将 SIFT 点舍弃不用。另外,求解刚性变换实际只需要 3 ~ 4 个点就够了,SIFT 特征过多反而浪费时间,因此将每幅影像上最多保存 1 000 个 SIFT 特征。

2)特征归并

通过特征提取,每一张影像上都提取出一系列特征点,而由于照片之间有重叠,以及扫描时可能同一区域经过了多次重复扫描,因此一组影像中有许多重复的特征,故可通过特征归并减少数据的冗余,提高后续特征匹配的效率。

SIFT 特征描述子(descriptor)具有 128 个方向,也就是说单个 SIFT 特征可以看做 128 维的向量,如果直接利用 SIFT 特征进行特征归并比较消耗资源(内存资源与计算资源),而且归并的效果不一定好。通过特征提取和特征映射,SIFT 特征点在点云数据中有一一对应的点(x,y,z)(为后续叙述方便称之为三维映射点)。三维映射点是从体元模型中得到的,因此处于统一的坐标系下,如果 SIFT 特征相同,则其对应的三维映射点的坐标应该十分接近,如图 4 所示。利用这一特性,可以利用三维映射点进行特征归并。

特征归并的过程其实就是邻近点聚类的过程。在欧式空间中何为最邻近,这里给出数学描述:给定一个多维空间 *RK*,*RK* 中的一个向量是一个样本点,这些样本点的有限集合称为样本集,给定样本集 E 和一个样本点 s',s'的最近邻就是任意样本点 $s\in E$ 满足 Nearest(E,s',s)。Nearest 定义如下:

$$\text{Nearest}(E,s',s)\Leftrightarrow\forall s''\in E\,|s'-s|\leqslant|s'-s''| \tag{2}$$

上面的公式中距离度量是欧式距离,即

$$|s'-s|=\sqrt{\sum_{i=1}^{k}(s_i-s'_i)^2} \tag{3}$$

其中:s_i 是向量 s 的第 i 个维度。由上式可见,搜索是一个线性过程,而由于数据量较大,逐点搜索肯定速度较慢,实验中可以采用 K-D 树或八叉树(Octree)[12,13]将样本集剖分,再进行邻近搜索,在本文中使用了成熟的代码库 FLANN[14],该代码库实现了基于 K-D 树的最邻近搜索,已在很多软件中得到广泛使用。

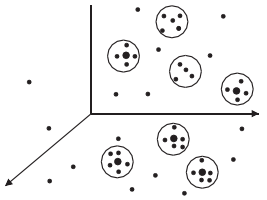


图 4　基于空间关系的特征归并

3)特征匹配与变换矩阵

当两幅图像的 SIFT 特征向量生成后,下一步采用关键点特征向量的欧式距离作为两幅图像中关键点的相似性判定度量。取图像 1 中的某个关键点,并找出其与图像 2 中欧式距离最近的前两个关键点,在这两个关键点中,如果最近的距离除以邻近的距离少于某个比例阈值,则接受这一对匹配点。降低这个比例阈值,SIFT 匹配点数目会减少,但更加稳定。另外,由于匹配时使用的是所有照片的 SIFT 特征,尽管已经归并了部分 SIFT 特征,两组相片中的 SIFT 特征仍然十分丰富,因此使用 GPU 实现的 SIFT 匹配算法提高了效率。

然而,由于 SIFT 匹配仍存在无匹配情况,因此通常利用 RANSAC[15]方法,通过 inlier 数据求解出合适的模型参数,基于该模型参数得到不适合该模型的 outlier,即异常点。

4)精匹配与数据拼接

通过 RANSAC 算法可以计算得到比较可靠的变换矩阵,但是利用该变换矩阵只能实现两组点云数据间的粗匹配,此时若用变换矩阵匹配两组点云数据不能得到完美的拼接效果,因此要在粗匹配基础上使用 ICP 算法实现点云数据间的精匹配。需要指出的是,参与精匹配的两组点云数据是体元生成的完整点云数据,相对于单张 RGB 影像对应的点云数据而言,完整点云数据更为丰富,使用 ICP 算法时更加可靠。通过 ICP 匹配可以计算得到一个新的矩阵,将其乘以原有的粗配准矩阵就可生成最终的精配准矩阵。将该矩阵应用于两组点云数据之上就可以实现数据的精确拼接。

3　实验结果与分析

3.1　实验环境

1)硬件环境

微软 Kinect 设备一台,包括数据转接线、USB 连接线等;计算机一台,处理器为 Intel® Core™ 2 Duo CPU E7400 2.80

GHz,内存 2 GB,硬盘 500 GB,GPU 型号为 NVIDIA GeForce GTX 560 Ti,显存频率为 4008 MHz,显存容量为 1 024 MB。

2)软件配置

操作系统 Microsoft Windows7 64 位;Kinect 驱动,PrimeSense 官方版本 OpenNI 框架;开发工具为 Microsoft Visual Studio 2010;开源库为 PCL(point cloud library)[16]。

3.2 实验结果

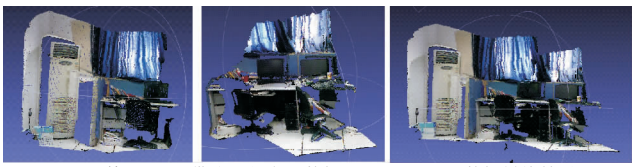
在进行多模型拼接实验时,至少需采集两组数据,包括关键帧数据集、相机位置参数、体元模型生成的点云数据。采集时需要注意,为重建出完整的模型,需要对扫描区域进行比较完整的扫描,当扫描区域有大片的平面区域时要格外小心,因为很可能会导致相机位置追踪失败。另外需要保证两组模型间有重叠区域这样才能进行数据匹配。

经过特征匹配后就可以使用 RANSAC 算法计算变换矩阵,利用变换矩阵实现两组点云数据的粗匹配,为展示数据拼接效果,使用点云数据可视化软件 Meshlab 将处理后的点云数据加载并显示,如图 5 所示。

(a) 第一、二组数据 (b) 拼接结果

图 5 多模型拼接粗匹配结果

在另一组实验中,本文在采集的原始点云数据基础之上添加准确映射的纹理信息,此时拼接的结果如图 6 所示。



(a) 第一、二组带纹理的点云数据 (b) 数据拼接结果

图 6 精匹配结果

实验结果表明,基于匹配的方法可以实现模型之间的准确拼接。与此同时,笔者发现数据拼接的结果受多方面因素的影响,一方面是场景自身的特性,如果场景中存在大量的重复性区域,匹配时可能会出错,尤其在场景中同时存在大片纹理稀疏区域时,这一问题更加明显;另一方面,在使用 RANSAC 和 ICP 算法时参数是否设置合理,参数的设置直接关系到矩阵求解的精度,并影响最终的模型拼接。

4 结束语

利用激光扫描仪获取的点云数据进行真实场景三维重建已有大量研究,然而应用于复杂的室内环境中存在极大限制。Kinect 作为一种新型深度传感器,可以快速获取大量的深度数据,而且体型小巧易于操作。因此,本文详细探讨了基于 Kinect 深度影像的数据融合方法,并将其应用于室内三维重建中。通过实验可以看到,基于体元技术的深度数据融合方法与传统的点云拼接技术有很多相似之处,并且它可以实时快速地重建出高精度三维模型,而不需要经过繁复的纹理映射就可重建出具有照片真实感的三维模型。本文解决了多模型数据拼接的问题,实现了大型室内场景的三维重建,对复杂的室内环境能精细地表达。通过实验分析可以看出,融合 Kinect 深度影像重建复杂室内场景能取得很好的效果,将会有广泛的应用前景。

参考文献:

[1] REMONDINO F. Heritage recording and 3D modeling with photogrammetry and 3D scanning[J]. Remote Sensing, 2011, 3(6): 1104-1138.

[2] BAUMGART B G. Winged-edge polyhedron representation[R]. Palo Alto, California: Computer Science Department, Stanford University, 1972.

[3] ELFES A. Using occupancy grids for mobile robot perception and navigation[J]. Computer, 1989, 22(6): 46-57.

[4] HOPPE H, DEROSE T, DUCHAMP T, *et al*. Surface reconstruction from unorganized points[C]//Proc of the 19th ACM Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1992: 71-78.

[5] CURLESS B, LEVOY M. A volumetric method for building complex models from range images[C]//Proc of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. 1996: 303-312.

[6] HENRY P, KRAININ M, HERBST E, *et al*. RGB-D mapping: using Kinect-style depth cameras for dense 3D modeling of indoor environments[J]. The International Journal of Robotics Research, 2012, 31(5): 647-663.

[7] HENRY P, KRAININ M, HERBST E, *et al*. RGB-D mapping: using depth cameras for dense 3D modeling of indoor environments[C]//Proc of International Symposium on Experimental Robotics (ISER). 2010.

[8] IZADI S, KIM D, HILIGES O, *et al*. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera[C]//Proc of the 24th Annual ACM Symposium on User Interface Software and Technology. 2011: 559-568.

[9] NEWCOMBE R A, DAVISON A J, IZADI S, *et al*. KinectFusion: real-time dense surface mapping and tracking[C]//Proc of the 10th IEEE International Symposium on Mixed and Augmented Reality. 2011: 127-136.

[10] LOW K L. Linear least-squares optimization for point-to-plane ICP surface registration[R]. [S. l.]: Department of Computer Science, University of North Carolina at Chapel Hill, 2004.

[11] BESL P J, MCKAY H D. A method for registration of 3-D shapes[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1992, 14(2): 239-256.

[12] KEIM D A. Efficient geometry-based similarity search of 3D spatial databases[C]//Proc of ACM SIGMOD International Conference on Management of Data. 1999: 419-430.

[13] 宋涛. 八叉树编码体数据的快速体绘制算法[J]. 计算机辅助设计与图形学学报, 2005, 17(9): 1990-1996.

[14] MARIUS M. FLANN, fast library for approximate nearest neighbors [EB/OL]. (2011). http://mloss.org/software/view/143/.

[15] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.

[16] RUSU R B, COUSINS S. 3D is here: point cloud library[C]//Proc of IEEE International Conference on Robotics and Automation. 2011.

[17] DES BOUVRIE B. Improving RGB-D indoor mapping with IMU data [D]. [S. l.]: Delft University of Technology, 2011.

[18] LOWE D. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.