基于深度相机的三维重建技术综述

摘要：三维重建是计算机图形学和计算机视觉领域的重要课题。近年来，消费级的深度相机的出现给三维重建带来了深刻的进步。本综述从基本数据结构出发，按照基于深度相机的三维重建的基本流程，详细分析了 RGB-D 三维重建的最新进展，并回顾了必要的相关工作。

关键词：深度相机，三维重建，体表达，面元

1. 引言

三维重建是计算机图形学和计算机视觉领域的重要研究方向，其核心内容是综合利用传感器和计算技术实现对三维物理世界的数字化表达，捕捉物体和场景高真实感的三维形状和外观，以在数字化空间模拟三维的交互与感知。

近年来，深度（或者称为 RGB-D）相机的出现，使得三维重建技术得到了飞速的发展。深度相机不仅价格低廉，体积小巧，并且能够以足够的分辨率和帧率捕捉像素级别的颜色和深度信息。正是由于这些特点，深度相机相较于那些更加昂贵的扫描设备在面向消费者的应用上具有巨大的优势。在其问世之初，研究者们就认识到利用深度相机进行三维重建的巨大潜力。帝国理工和微软研究院在 2011 年提出的 KinectFusion[IKH11，NIH*11]开启了利用深度相机进行三维重建的序幕。KinectFusion 系统能够利用深度相机获得的彩色和深度图像实时地生成具有高清细节的 3D 模型，这得到了三维重建领域研究人员的高度关注。近年来计算机图形学和计算机视觉领域的研究者在基于深度相机进行三维重建的算法和系统上不断创新，取得了丰富的研究成果。

基于深入的文献调研，本综述将详细回顾和比较基于深度相机的三维重建的最新方法。本文首先在第 2 节给出了三维重建的两种基本数据结构，着重在第 3 节介绍了基于深度相机的几何重建的研究进展，最后在第 4 节对全文进行了总结，给出了未来研究方向的展望。

2. 基本数据结构

2.1 体表达

体表达的概念首次提出是在 1996 年，Curless 和 Levoy [CL96]介绍用符号距离函数（SDF）来表达模型的方法。他们将物体表面定义为零，自由空间（即物体外部）为一个正值，值的大小随着其到最近表面的距离的增大而增大，而占有空间（即物体内部）则是一个

相似的负值。这些函数值被存储在一个栅格立方体中。在执行算法之前，必须定义立方体的体素大小以及空间范围。其他数据，如颜色，通常存储在每个体素属性中。由于只有靠近实际表面的体素是重要的，研究者们通常采用截断符号距离函数（TSDF）来表达模型。由于TSDF 值隐式地表现了物体的表面，在相机位姿估计前需要用光线投射法等方法提取出物体表面。融合步骤通过简单的加权平均来实现，这种对多距离样本的时间积分对于消除深度相机的噪声十分有效。

常规的体素网格在内存消耗方面非常低效，并且受限于预定义的立方体和分辨率。在实时场景重构的背景下，大多数方法都严重依赖于现代 GPU 的处理能力，体素网格的空间范围和分辨率通常受到 GPU 内存的限制。为了支持更大的空间范围，研究者们已经提出了多种方法来提高基于体表达的算法的内存效率。为了防止在预定义的立方体之外获取的深度图像而造成的数据丢失，Whelan 等人[WKF12]提出了体素网格动态移位的方法，使体素网格跟随深度相机的运动。该方法每当体素网格移动时，提取出当前体素网格之外的部分，并单独存储，添加其相机位姿到全局位姿图。虽然这可以实现更大的扫描量，但它需要大量的外存使用，并且已经扫描的表面不能被任意重新访问。Henry 等人[RV12]也提出了类似的思想。

体素层次结构，比如八叉树，是一种能够高效存储物体表面的方式。八叉树可以根据场景的复杂度自适应地分割场景空间，有效地利用计算机的内存[FPRJ00]。虽然其定义简单，但由于节点的稀疏性，在利用 GPU 的并行性上很有难度。Sun 等人[SZKS08]建立了一个只有叶子节点的八叉树来存储体素数据，并基于八叉树加速跟踪。Zhou 等人[ZGHG11]利用 GPU 构建了一个完整的八叉树结构，以可交互的速率对拥有 30 万个顶点的场景进行泊松重建。Zeng 等人［ZZZL13］扩展了这种数据结构，为 KinectFusion 实现了一个 9 到 10 级的八叉树，并将重建结果扩展到一个中等规模的办公室。Chen 等人[CBI13]提出了类似的 3 级层次结构，同样具有优秀的分辨率。Steinbrücker 等人[SSC14] 以多分辨率的数据结构表示场景，以便在 CPU 上进行实时累积，以大概 1Hz 的频率输出实时的网格模型。Henry 等人[HFBM13]提出了碎片(patch)体表达。这种方法将整个分割成许多具有任意大小和分辨率的碎片体块。碎片体块在空间上被组织成一个位姿图，为了得到一个全局一致的模型，需要对这个位姿图进行优化。这个方法具有交互性，但不是完全实时的。

Nießner 等人[NZIS13]提出了体素哈希结构。在这种方法中，体素的空间位置所对应的索引存储在线性化的空间哈希表中，并通过空间哈希函数进行寻址。由于内存中只保存那些包含空间信息的体素，这种策略大大减少了内存的消耗，在理论上允许设定一定大小和分辨

率的体素块来表示无限大的空间。与体素层次结构相比，哈希结构在数据插入和访问上具有巨大优势，其时间复杂度均为 O(1)。但哈希函数不可避免地将多个不同的空间体素块映射到哈希表中的同一个位置。Kahler 等人[KPR15]采用不同的哈希方法来减少哈希冲突的数量。基于哈希的三维重建拥有内存需求小，计算高效的优势，使其甚至适用于移动设备，比如被应用在谷歌的 Tango 上[DKSX17]。

2.2 面元

面元是另一种重要的模型表达方式，最初被用作模型渲染的基元[PZVG00]。，一个面元通常包含下面几个属性：空间坐标 $p \in \mathbb{R}^3$，法向量 $n \in \mathbb{R}^3$，颜色信息 $c \in \mathbb{N}^3$，权重（或者置信度）$w \in \mathbb{R}$，半径 $r \in \mathbb{R}$，时间戳 t。面元的权重用于融合时的加权平均以及对稳定点和非稳点的判断，一般用下式初始化：$w = e^{-\gamma^2/2\sigma^2}$。其中$\gamma$是当前深度测量到相机中心的归一化径向距离，$\sigma$通常取 0.6[KLL*13]。面片的半径通过场景表面离相机光心的距离的求得，距离越大，面片的半径越大。相较于体表达，基于面元的三维重建对于从深度相机获得的输入表达得更简洁、更直观，无需在不同的表示之间来回转换。与体表达类似，八叉树可以被用于改进基于面元的数据结构[SB12,SB14]。

3. 基于深度相机的几何重建

3.1 深度图预处理

深度相机获得的深度图像具有一定的噪声。大多数情况下，我们使用双边滤波[TM98]来降低深度图像的噪声。降噪之后，[IKH11]应用均值降采样获得了一个 3 层的深度图金字塔，这是为了在下一步从粗到细地估计相机位姿。对于降噪后的深度图像，根据深度相机的内参，可以投影得到每个像素点在相机坐标系的三维坐标，称为顶点图。每个顶点的法向量可以通过该顶点与相邻顶点的叉乘获得。对于面元表示，还需要计算每个点的半径，用来表示给定点周围的局部表面积，同时最小化相邻点之间的可见孔，如图 2 所示。[SGKD14]用下式计算顶点半径：$r = \sqrt{2} * d / f$，d 代表顶点的深度值，f 表示深度相机的焦距。

3.2 相机位姿估计

通常，某一时刻的 6 自由度的相机位姿用一个刚性变换矩阵 T_g 来表示。这个变换矩阵

将相机坐标系映射到世界坐标系，该时刻的某一点 p 在相机坐标系下的三维坐标 p_k 和世界坐标系下的三维坐标 p_g 通过下式转换：$p_g = T_g * p_k$。估计相机位姿是将当前帧的深度图像融合到全局模型的前提。

3.2.1 体表达

由于体表达用 SDF 值隐式地储存场景的表面，在估计相机位姿时需要根据前一帧的相机位姿提取出世界坐标系下已重建的模型的表面。表面预测通常采用了光线透射法[PSL*98]。

3.2.1.1 跟踪目标

对于相机跟踪，迭代最近邻算法(ICP)是一种重要的方法。早期，迭代最近邻算法被应用于三维形体注册中[BM92，CM91]，通过在相邻的两帧之间寻找匹配点，计算刚性变换，使这些点对之间的距离平方和最小，并迭代，直到满足某种收敛准则。采用这种帧到帧的策略的一个严重的问题是，每两帧的注册都会产生误差，这种误差会随着扫描进行不断地累积。

为了减轻这个问题，近年来的基于深度相机的重建方法广泛地采用了帧到模型的相机跟踪方法[NIH*11，WKF12，NZIS13]。虽然帧到模型的跟踪显著降低了每一帧的跟踪漂移，但并不能完全解决误差积累的问题，因为跟踪误差仍然会随着时间累积。而累积的漂移最终会导致重建表面在轨迹闭环处的不重合。因而，研究学者引进了全局位姿优化的思想。Zhou 等人[ZMK13] 使用兴趣点来保存局部细节，并结合全局位姿优化来均匀分布场景中的注册误差。Zhou 等人[ZK13]将视频序列分割成片段，使用帧到模型的集成，从每个片段中重建局部精确的场景片段，在交叠的片段之间建立密集的对应关系，并优化一个全局目标函数来对齐片段。优化后的图像可以对图像碎片进行细微的变形，从而纠正输入图像中低频失真引起的不一致性。这两个场景重建方法都是离线的。Dai 等人[DNZ*17] 基于在线的束调整和表面重新集成，在实时的帧率下实现了全局一致的重建。

3.2.1.2 数据关联

不管是基于哪种策略的相机跟踪，寻找对应点都是必不可少的一步。对应点对的集合带入到待优化的目标函数中，从而计算出相机位姿的最优解。寻找对应点按照所使用的点对可分为稀疏的方法和稠密的方法。稀疏的方法即只对特征点找对应，而稠密的方法则是对当前帧的所有点寻找对应。

对于特征点的提取和匹配，SIFT 算法[LOWE04]是一个广受欢迎的选择。SURF 算法[BETV08]对 SIFT 算法进行了改进，提高了检测特征点的速度。ORB 特征描述符是可以代替 SIFT 和 SURF 的一种选择，它的速度更快，相应的稳定性也较差。Endres 等人[EHE*12]在寻找特征点时分别采用了这三种特征描述符，并对它们的效果进行了比较。Whelan 等人

[WKF*12]使用基于特征点的视觉快速测程法(FOVIS)代替了 KinectFusion 中基于迭代最近邻的方法，并且在闭环检测时运用了 SURF 特征描述符。Zhou 等人[ZK15]提出了基于物体轮廓的相机跟踪，引入了轮廓特征的对应点，使跟踪更稳定。最近的 BundleFusion [DNZ*17]采用 SIFT 特征进行粗配准，再用稠密的方法进行精配准。

对于稠密的方法，传统的寻找对应点的方法十分耗时。投影数据关联算法[BL95]极大地加速了这一过程。这种策略根据相机位姿和相机内参，将输入点的三维坐标投影到目标深度图上的一个像素，通过这个像素的邻域内的搜索，得出最佳的对应点。对应点的误差有很多度量方式，比如点到点的度量[BM92]，点到面的度量[CM91]。相比于点到点的度量方式，点到面的度量方式收敛速度更快，许多方法都采用点到面的度量方式。除了将几何上的差异作为约束项，许多方法将对应点之间光学上的差异（即颜色差异）作为约束项，比如 [SKC13, DNZ*17]。

3.2.2 面元

Henry 等人[HKH12]使用稀疏特征点来匹配当前帧与之前的观测值，同时也考虑了空间因素，结合使用稠密的点云注册。他们在特征点匹配上，采用了 SIFT 特征描述符，采用 RANSAC 算法[FB81]确定拥有一致的刚性变换的特征对子集。对于回环检测，Henry 等人提出了关键帧的概念，每当累计的相机位姿大于一个阈值时产生一个关键帧。回环检测只在出现关键帧时进行，使用 RANSAC 算法对当前关键帧与先前的关键帧进行特征点匹配。检测到回环后，有两种全局优化的策略可以选择。一是位姿图优化，即用位姿图结构表示帧与帧之间的约束，帧与帧之间的边对应着几何约束。二是稀疏的束调整，即全局最小化在所有帧上使用 SBA 匹配的特征点的再投影误差[TMHF00,LA09, K10]。

Whelan 等人[WLS*15]结合使用深度和颜色信息来对每个输入帧进行全局位姿估计。用于注册的深度图和颜色图的预测通过对全局模型采用表面溅射算法[PZVG00]得到。他们将回环问题分为全局回环和局部回环。全局回环检测使用随机蕨编码的方式[GSCI15]。对于回环优化，他们采用了基于嵌入式形变图[SSP07]的空间形变方法。形变图由一组节点和边缘组成，这些节点和边缘分布在要形变的模型中。每个面元受一组节点影响，影响权重与面元到节点的距离成反比。

Lefloch 等人[LKS*17]将曲率信息作为其独立的表面属性进行实时重建。他们不仅在寻找对应点时考虑了输入图和全局模型的曲率，而且在对目标函数介绍了基于曲率的加权方案。这显著地提高了 ICP 收敛的鲁棒性，从而使位姿估计的漂移最小化。

3.3 融合

3.3.1 体表达

基于体表达的三维重建的融合，首先需要计算输入的深度图所对应的符号距离函数值(SDF)。相对于计算真正的离散的符号距离函数，通常计算投影的截断符号距离函数(TSDF)，因为它更容易计算，更适于并行。输入的深度图对应的 TSDF 值通过加权平均的方式融合到全局模型的 TSDF。

3.3.2 面元

在估计了当前输入帧的相机姿态之后，将每个顶点以及相关的法线和半径集成到全局模型中。深度图的融合可以分为三个步骤：数据关联，加权平均，移除无效点[KLL*13]。通过将全局模型投影到当前相机视角的像平面寻找对应点。由于相近的模型点可能投影到同一个像素上，采用上采样的方法提高精度。如果找到了对应点，则使用加权平均将最可靠的点与新的点估计合并。如果没有找到可靠的对应点，则将新的点估计作为不稳定点添加到全局模型中。随着时间的推移，全局模型会被清理，以消除由于可见性和时间限制而产生的异常值。

4. 小结

在过去几年里，由于消费级的深度相机的广泛使用，使用普通深度相机进行在线和交互式的三维建模有了很大的发展。目前的开发覆盖了整个重建管道，并带来了创新的各个层次和中间步骤，从深度相机硬件到高层次的应用，涉及到日常生活的方方面面。

尽管取得了很大进展，基于深度相机的三维建模仍然面临地诸多挑战。比如：如何更好地重建动态的场景；开发更通用的场景重建方法，即在单一方法中涵盖更广泛的应用场景；RGB-D 传感器已经在移动设备和智能手机上出现，如何在这些设备上进行重建。

参考文献

[BM92] BESL P. J., MCKAY N. D.: A method for registration of 3-d shapes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, TPAMI, 14*, 2 (1992), 239–256.

[BETV08] Bay H., Ess A., Tuytelaars T., Van Gool L.: Surf: Speeded Up Robust Features. In *Computer Vision and Image Understanding 10* (2008)*, 346–359*

[BL95] Blais G., Levine M. D.: Registering multiview range data to create 3D computer objects.In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *17*, 8(1995),pp. 820–824.

[CM91] CHEN Y., MEDIONI G.: Object modeling by registration of multiple range images. In *IEEE International Conference on Robotics and Automation, ICRA*, 1991, pp. 2724–2729 vol.3.

[CBI13] Chen J., Bautembach D., Izadi S.: Scalable real-time volumetric surface reconstruction. In *ACM Trans. on Graphics (Proc. SIGGRAPH) 32*, 4 (2013), 113:1–113:16.

[CL96] Curless B., Levoy M.: A volumetric method for building complex odels from range images. In *Proc. Comp. Graph. & Interact. Techn.* (1996), pp. 303–312.

[DKSX17] Dryanovski, I., Klingensmith, M., Srinivasa, S.S., Xiao, J.: Large-scale, real-time 3D scene reconstruction on a mobile device. In *Auton. Robots 41*, 6 (2017), pp. 1423–1445

[DNZ*17] DAI A., NIESSNER M., ZOLLHÖFER M., IZADI S., THEOBALT C.: Bundlefusion: real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. In *ACM Transactions on Graphics, TOG, 36*, 4 (2017).

[EHE*12] Endres F., Hess J., Engelhard N., Sturm J., Cremers D., and Burgard W.: An evaluation of the RGB-D SLAM system. In *Proceedings of the IEEE Int. Conf. on Robotics and Automation (ICRA)* (2012).

[FB81] Fischler M., Bolles R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24, 1981, pp. 381–395.

[FPRJ00] Frisken S.F., Perry, R.N., Rockwood, A.P., Jones, T.R.: Adaptively sampled distance fields: a general representation of shape for computer graphics. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH* (2000), pp.249–254.

[GSCI15] Glocker B., Shotton J., Criminisi A., and Izadi S.: Real-Time RGB-D Camera Relocalization via Randomized Ferns for Keyframe Encoding. *IEEE Transactions on Visualization and Computer Graphics*, 21(5), 2015, pp. 571–583.

[HKH12] HENRY P., KRAININ M., HERBST E., REN X., FOX D.: RGB-D mapping: Using depth cameras for dense 3d modeling of indoor environments. In *International Symposium on Experimental Robotics, ISER* (2010), pp. 477–491.

[IKH11] Izadi S.,Kim D., Hilliges O.,Molyneaux D., Newcombe R.,Kohli P., Shotton J., Hodges S., Freeman D., Davison A., Fitzgibbon A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proc. ACM Symp. User Interface Softw. &Tech.,* 2011, pp. 559–568.

[KLL*13] Keller M., Lefloch D., Lambers M., Izadi S., Weyrich T., Kolb A.: Real-time 3D reconstruction in dynamic scenes using pointbased fusion. In *Proc. Int. Conf. 3D Vision (3DV)* (2013), IEEE Computer Society, pp. 1–8.

[KPR15] Kahler O., Prisacariu V. A., Ren C. Y., Sun X., Torr P., urray D.: Very high frame rate volumetric integration of depth images n mobile devices. In *IEEE Trans. on Visualization and Computer Graphics 21*, 11 (2015), 1241–1250.

[LA09] Lourakis M., Argyros A.: SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software* 36, 2009, pp. 1–30.

[LKS*17] Lefloch, D., Kluge, M., Sarbolandi, H., Weyrich, T., Kolb, A.: Comprehensive Use of Curvature for Robust and Accurate Online Surface Reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* (2017), pp. 1.

[LOWE04] Lowe D.: Distinctive image features from scale-invariant keypoints, cascade filtering approach. In *International Journal of Computer Vision60* (2004), pp. 91–110.

[K10] Konolige K.: Sparse sparse bundle adjustment. In *Proceedings of the British Machine Vision Conference (BMVC),* 2010.

[NIH*11] Newcombe R. A., Izadi S., Hilliges O., Molyneaux D., Kim D., Davison A. J., Kohi P., Shotton J., Hodges S., Fitzgibbon A.: Kinectfusion: Real-time dense surface mapping and tracking. In *International Symposium on Mixed and Augmented Reality, ISMAR*, (2011), pp. 127–136.

[NZIS13] Nieÿner M., Zollhofer M., Izadi S., Stamminger M.: Realtime 3D reconstruction at scale using voxel hashing.In *ACM Trans. On Graphics 32*, 6 (2013), 169.

[PSL*98] Parker S., Shirley P., Livnat Y., Hansen C., and Sloan P.: Interactive ray tracing for isosurface

rendering. In *Proceedings of Visualization*, 1998.

[PZVG00] Pfister H., Zwicker M., Van B. J., Gross M.: Surfels: Surface elements as rendering primitives. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2000), SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., pp. 335–342.

[RV12] ROTH H., VONA M.: Moving Volume KinectFusion. In *British Machine Vision Conference, BMVC*, 2012, pp. 1–11.

[SB12] Stuckler J., Behnke S.: Integrating depth and color cues for dense multi-resolution scene mapping using rgb-d cameras. In *IEEE Intl. Conf. on Multisensor Fusion and Information Integration (MFI)*, 2012.

[SB14] Stückler J., S. Behnke. Multi-resolution surfel maps for efficient dense 3d modeling and tracking. *Journal of Visual Communication and Image Representation*, 25(1), 2014, pp. 137–147.

[SGKD14] Salas-Moreno R., Glocker B., Kelly P. H. J., and Davison A. J.: Dense Planar SLAM. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2014.

[SKC13] Steinbrucker F., Kerl C., Cremers D.: Large-scale multiresolution surface reconstruction from rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 3264–3271.

[SSC14] Steinbrucker F., Sturm J., Cremers D.: Volumetric 3D mapping in real-time on a cpu. In *Proc. IEEE Int. Conf. Robotics and Automation* (2014), IEEE, pp. 2021–2028.

[SSP07] Sumner, R.W., Schmid, J., Pauly, M.: Embedded deformation for shape manipulation. In *ACM TOG 26*, 3, 2007.

[SZKS08] Sun, X., Zhou, K., Stollnitz, E., Shi, J., Guo, B.: Interactive relighting of dynamic refractive objects. In *ACM Trans. Graph. 27*,3 (2008), 35.

[TM98] Tomasi C., Manduchi R.: Bilateral filtering for gray and color images. In *Proc. IEEE Int. Conf. on Computer Vision* (1998), pp. 839– 846.

[TMHF00] Triggs B., McLauchlan P., Hartley R. and Fitzgibbon A.: Bundle adjustment—a modern synthesis. *Vision Algorithms: Theory and Practice* 1883, 2000, pp. 153–177.

[WKF*12] Whelan T., Kaess M., Fallon M., Johannsson H., Leonard J., McDonald J.: Kintinuous: Spatially extended kinectfusion. In *Proc. RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras* (2012).

[WLS*15] WHELAN T., LEUTENEGGER S., SALAS-MORENO R. F., GLOCKER B., DAVISON A. J.: Elasticfusion: Dense SLAM without A pose graph. In Robotics: Science and Systems XI, Sapienza University of Rome, 2015.

[ZGHG11] Zhou K., Gong M., Huang X., Guo B.: *Data-parallel octrees for surface reconstruction. IEEE Transactions on Visualization and Computer Graphics (TVCG) 17*, 5 (2011), pp. 669–681.

[ZMK13] Zhou Q.-Y., Miller S., Koltun V.: Elastic fragments for dense scene reconstruction. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2013), pp. 473–480.

[ZK13] Zhou Q.-Y., Koltun V.: Dense scene reconstruction with points of interest. In *ACM Trans. on Graphics 32*, 4 (2013), pp. 112.

[ZK15] Zhou Q.-Y., Koltun V.: Depth camera tracking with contour cues. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 632–638.

[ZZZL13] Zeng M., Zhao F., Zheng J., Liu X.: Octree-based fusion for realtime 3D reconstruction. *Graphical Models 75*, 3 (2013), 126–136.