文章编号：2095-2163(2020)07-0049-04 中图分类号：TP391.41 文献标志码：A

单目三维重建技术综述

包晓敏，白 晨
（浙江理工大学 信息学院，杭州 310018）

摘 要：计算机视觉领域的三维重建主要指从二维图像还原出物体的三维结构。本文主要介绍了单目三维重建技术，即使用单个摄像机所成图像，寻找图像的对应特征，根据这些特征的变化建模，逆向推导出物体的三维信息。本文主要从单目三维重建较为典型的单目三维重建算法包括明暗法、光度立体视觉法、轮廓法等来展现这种对应特征的建模问题，并介绍了这些算法的适用情况及不足。

关键词：三维重建；单目；对应特征；建模

Overview of monocular 3D reconstruction techniques

BAO Xiaomin, BAI Chen
(Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 In the field of computer vision, three-dimensional reconstruction mainly refers to restoring the three-dimensional structure of objects from two-dimensional images. This paper mainly introduces the technique of monocular 3D reconstruction, that is, using the image produced by a single camera, looking for the corresponding features of the image, modeling based on the changes of these features, and reversely deducing the 3D information of the object. In this paper, the typical monocular 3D reconstruction algorithms including light and shade method, photometric stereo method, contour method, etc. are mainly used to show the modeling problems of corresponding features, and the application and shortcomings of these algorithms are introduced.

【Key words】 3D reconstruction; monocular; corresponding features; modeling

0 引 言

计算机视觉中的三维重建是指从二维图像中还原出物体三维结构的过程，是计算机视觉中研究价值较高的一个方向。三维重建有很高的实用价值，广泛应用于医疗影像、航空航天、文物保护、遥测、机器人行走和无人驾驶等领域。

三维重建技术按照摄像机数目分类，可以分为：单目、双目和多目。单目重建技术是指用单个摄像头拍摄的图像进行重建，分析单张图像或者序列图像的特征变化，建立空间结构与特征的关系模型，重构物体三维结构。相比于双目和多目，单目重建更节省成本，重建难度也较高。目前的单目重建方法有：明暗度法、光度立体视觉法、纹理法、轮廓法、焦距法、运动法及深度学习法等。

1 明暗度法

明暗度法，即明暗度恢复形状法（shape from shading，SFS），主要是根据一定光照强度下，不同距离，不同曲率表面成像的明暗度差异进行建模。Horn 于 1970 年首次提出此方法。最初的 SFS 是通过建立图像对应位置像素点明暗度与光照方向，物体反射率及表面梯度之间的关系模型，来恢复深度信息。但是这种 SFS 是一个欠约束问题，需要增加约束才能求解[1]，所以传统的 SFS 加入了 3 个假设条件[2]：

（1）反射模型为朗伯体表面模型。

（2）光源为无限远处光源或平行光。

（3）成像关系为正交投影。

这些假设使求解容易，但也造成了算法的局限。一方面朗伯模型是一种理想化模型，真实物体并不满足这种模型；另一方面正交投影的假设也使得求解变成一个病态问题。针对这两点局限，Yang Lei 使用混合反射模型代替朗伯模型，建立哈密顿-雅克比方程关系模型，解决了镜面反射问题，扩大了明暗度法的适用范围，同时采用数值方法计算哈密顿-雅克比方程，加快了处理的速度，满足了实时重建的需求[3]。为了解决由光源方向不准确而导致的

基金项目：浙江省重点研发计划项目（2018C01133）。

作者简介：包晓敏（1965-），女，硕士，教授，主要研究方向：计算机软件及计算机应用、自动化技术、图像处理等；白 晨（1995-），女，硕士研究生，主要研究方向：图像处理。

通讯作者：包晓敏 Email：1316831475@qq.com

收稿日期：2020-05-27

三维信息误差较大的问题,杨志明等建立了基于径向基函数神经网络的反射模型,代替了理想朗伯体表面反射模型[4]。

明暗度法能够根据单幅图像得到较为精确的三维模型,但是对光照要求很严格,难以在室外场景应用。

2 光度立体视觉法

光度立体视觉法依据像素灰度值与物体表面法线、反射率及光源向量的关系模型,通过解算物体表面在光照下像素变化的矩阵来得到法线信息,再通过表面法线的积分来得到物体的三维信息。与明暗度法不同的是光度立体视觉方法一般使用多幅图像进行计算,这样就给对应像素位置添加了约束,这也造成了矩阵计算中的计算量和精度问题。

光度立体视觉的主要问题是对物体表面在光照下的像素变化矩阵的精确求解。吴仑等通过鲁棒的主成分,分析得到低秩矩阵和表面向量场,再使用表面重建算法从向量场恢复物体形状,提高了噪声情况下物体表面的重建精度[5];Di Xu 针对光度立体视觉算法计算量大的问题提出了一种两步优化算法来恢复表面,即先用阴影信息优化面法线,再通过面法线更新顶点[6],与现有算法相比,显著减少了运行时间,并能够产生更加密集的网络。Hashimoto 通过图像分解得到相机内参,结合物体形状的近似估计信息,解决了未标定情况下的光度立体视觉图像的三维重建问题[7];付琳等提出一种基于共位图像的逆向反射模型,实现了从图像像素值到法向量与入射光线内积的精确映射,仅使用一张共位图像和一张多光谱条件的 RGB 图像即可实现高精度的光度立体视觉[8],大大缩短了拍摄所需时间,同时使用神经网络拟合近场光度立体视觉模型中的映射关系,提高了算法的鲁棒性;陈明汉提出了一种基于深度学习多尺度卷积架构的光度立体视觉算法,实现了对非透明材质表面在任意光照条件下的高精度向保留精细三维形貌恢复能量恢复[9],算法在力的同时,极大地提高了针对非朗伯表面的适应性,为光度技术的广泛应用提供了有力的技术支持。

光度立体视觉算法是光学测量领域的重要方法之一,可以对物体表面纹理进行大范围、超精度的三维重建。缺点是不能处理镜面反射及室外场景的三维重建。

3 轮廓法

轮廓法即轮廓恢复形状法(shape from silhouettes, SFS),Baumgart 于 1974 年提出,通过延长多个角度的相机光心到图像中物体轮廓的线段,产生的重叠区域重建物体的三维模型[10]。这种方法主要是通过不同角度成像的轮廓划分空间区域,图像越多,划分的物体的区域就越细致,相应的计算量也越大。

轮廓法的原理相对简单,一般用于静态物体的三维重建。算法使用的图像一般要求是环绕物体的几个角度拍摄,并且要求已知各个图像对应的投影矩阵。轮廓法建立的模型较为粗糙,角度越丰富,对重叠区域的限制越多,重建效果越精细,因此轮廓法需要大量的图像进行细化,计算量大。Potmesil 在 1987 年提出使用八叉树来加速 SFS 的计算[11];Keith Forbes 加入两个镜子来代替摄像机并进行自校准,简化了算法的操作[12];Gloria 针对图像轮廓受噪声,背景扣除误差和遮挡影响而不均匀的现象,把轮廓和形状之间的误差作为能量,轮廓不均匀问题转化为能量最小化问题,从而恢复了表面形状[13]。

轮廓法的优点是原理简单,只使用了物体的轮廓信息,没用充分运用图像的信息,重建粗糙,所需的图片数量大。

4 纹理法

纹理法即纹理恢复形状法(shape from texture, SFT),由于物体的纹理是固定的,成像过程中物体固定的纹理元会由于相机与物体的距离和方向的变化而变化,根据这些变化可以反求物体的深度信息。

Kanatani 根据透视投影导致的纹理变形和图像坐标表示的“第一定律形式”,推导出用于确定平面和曲面的表面形状的方程式,并给出了求解这些方程的数值解法[14];Blake 和 Marinos 证明了纹理平面取向的统计估计是反投影问题,并应用“第二力矩反馈”得到了收敛的结果[15];Bostein 针对倾斜的纹理表面难以识别纹理元素的问题,使用基于高斯拉普拉斯标度空间的多尺度区域检测器构建候选文理,搜索图像确定图像的真实纹理[16];刘军等根据车道的平行车道的纹理特征,寻找消隐点,进而得到了图像中物体的深度信息[17]。

纹理法对图像的质量和物体的纹理要求都很高,需了解并识别成像中纹理的畸变,应用范围窄。而对于纹理特征明显的单张图像,使用纹理法则较为简单。

5 焦距法

焦距法即焦点恢复形状法(shape from focus, SFF),根据物体与相机距离等于焦距时成像最清晰的原理,建立焦距与图像清晰度的关系模型,调整焦距可以得到清晰度不同的图像,通过分析这些图像

的清晰度来计算深度信息。该方法于 1997 年由 Rajagoplan 首次提出。焦距法的主要问题是对聚焦量的准确计算和噪声对清晰度判断的影响。Sahay 考虑了焦距变化造成的视差效应，得到了更为准确的结果[18]。针对焦距法使用固定窗口的局部平均来增强初始聚焦量时，大窗口容易造成物体形状过度平滑，小窗口无法有效抑制噪声的问题，Mahmood 采用了迭代 3D 各向异性非线性扩散滤波器(ANDF)，在保留边缘的同时抑制了噪声，获得了不错的效果[19]。焦距法是一种重要的微观测量技术，在生活中应用较广，如元件磨损测量。

6　运动法

运动法即运动估计结构法(structure from motion，SFM)，首先进行特征提取和特征的匹配，然后根据特征点匹配对和对极几何原理得到相机的基础矩阵，由基础矩阵的奇异分解和三角形法得到特征点的三维坐标，最后使用 bundle adjustment(BA)最小化重投影误差，估计出相机的参数。SFM 按照图片加入方式不同可分为：增量式，全局式和混合式。增量式 SFM 一般以第一幅图像的相机姿态作为世界坐标，首先对两幅图像进行特征点的匹配、基础矩阵分解、根据初始相机参数和三角形法得到三维坐标，再在前两幅图像的基础上增加第三幅图像，将第二幅与第三幅进行特征点的匹配，并根据第二幅图像已知特征点的三维坐标和第二幅图像与第三幅图像的基础矩阵，得到第二幅图像与第三幅图像的旋转矩阵，由第二幅图像与第三幅图像的匹配点的三维坐标，得到第三幅图像的位姿，之后便可以计算第三幅图像的其余特征点在世界坐标中的三维信息，最后进行点云的融合，以此类推地加入图片，完善物体的三维信息。

增量式 SFM 是目前应用最广的方法，也是最早提出的方法，大多数算法在此基础上进行改进。Snavely 于 2007 年针对网络图像恢复三维结构的问题提出了 SFM 算法，算法采用 SIFT 特征点提取，通过五点法，三角化和 BA 计算两张视图的初始模型，再以增量的方式增加图片，三角化新的特征点轨迹[20]。算法已经具有完整的 SFM 结构，但是由于五点法对外点敏感，算法的准确性不高，且运行效率很低，难以在线应用。针对增量式相机模型估计严重依赖 BA 的情况，Moulon 提出一种基于逆模型估计的自适应 SFM 算法，在筛选内点时设计了最小化 NFA(Number of Falses alarm)的 AC-RANSAC，较传统的 RANSAC 算法选择内点的能力更好[21]。ChangchangWu 针对增量式 SFM 计算量大，效率低的问题，从三方面对算法做了改进，在特征提取上，使用 preemptive feature matching 选择特征进行匹配，仅对前 h 特征进行匹配；在计算过程中，引入了 precondtioned conjugate gradient(PCG)对 BA 进行优化，大大减少了算法的时间复杂度；对于增量式 SFM 累积误差造成的漂移问题提出了 re-triangulation[22]。改进算法虽然显著减少了 SFM 的运行时间，但对大场景的处理效果仍有限。Schonberger 从五点对 SFM 进行了改进[23]：

(1)使用一个多模态的几何校验策略来增强场景图中的几何关系。

(2)根据图像中点的数量对图像进行评分，选择最多可视化三角点的图像作为输入。

(3)利用传递对应性提升三角化的完整度和精度，解决稀疏图像集特征点不足的问题。

(4)只在点云数达到一定规模时再进行点云化，有效减少算法运行时耗。

(5)从无序图像集中寻找相似的视图，以发掘潜在的约束关系，提高 BA 效率。

根据增量式 SFM 算法误差累计造成的漂移问题，Cui 提出了 BSFM，即批量选择图片进行 SFM 和 BA，最后通过滤波器进行筛选[24]。

全局式 SFM 提取所有图片的特征点，并对所有图片进行相互的匹配，计算本质矩阵，根据本质矩阵奇异分解得到相机之间的旋转矩阵，以此建立整体的视角图，根据图片两两之间的匹配点数及匹配误差，从整体视角图中建立全局坐标系，使用全局坐标系获得图片在全局坐标系下的平移量和图片特征点的三维坐标，最后使用光束平差法(BA)进行三维重建。

增量式 SFM 研究较早，在重建精度和鲁棒性上效果较好，但在大规模场景下会存在漂移问题，主要是由于误差的累积导致场景不闭合，另外使用 BA 算法十分耗时且会产生不收敛的现象。全局式 SFM 只在最后执行 BA，效率较增量式 SFM 高，但对特征点匹配的准确度和精度要求很高。

7　深度学习法

深度学习对二维图像的三维结构重建过程可以理解为对一个函数的拟合，自变量是像素信息，因变量是图像的深度信息。David Eigen 设计了深度估计的 CNN 网络，网络分为两层，分别是全局预测网络和局部优化网络，训练时先对全局网络进行训练，再固定前者训练后者[25]。通过将全局信息与局部

信息结合,算法能够预测深度信息,但是相比于 ground-truth 较为模糊。2015 年,David Eigen 又提出了一个多任务的网络,这个网络可以预测深度、表面法向量和语义标签[26]。框架由三个网络构成,三个网络成串联结构;WeiFeng Chen 设计了两流式深度估计网络,上层网络估计图像的深度,下层网络估计深度的梯度[27],都采用了 VGG-16 网络结构作为图像解析模块,之后是分别联通了一个特征融合模块和细化模块,最后深度融合模块将深度和深度梯度进行融合得到最后的深度估计图,算法在 NYU Depth v2 上测试,得到很好的效果;Yue Luo 创造性地设计了生成右视图的网络,采用立体匹配的方法恢复图像的深度信息[28]。深度学习法不同于传统方法,没有固定的公式,是在设计好网络的结构和代价函数之后,让网络自己根据数据集提供的输入与结果来调整自己的参数,自己寻找规律。深度学习法的适用性依赖于数据集的丰富程度,网络结构的合理性。

8 结束语

单目三维重建高效利用了图像的信息,通过分析成像灰度值,纹理,轮廓,清晰度,位置等与深度的关系,建立对应的模型,再根据从图像中提取的特征信息值逆向求解出深度信息。在这一过程中,模型完善,噪声去除,精度和计算速度的提升,图像对应点定位的准确度的提高等一直是研究的重点。

参考文献

[1] Peter, N, Belhumeur, et al. The Bas-Relief Ambiguity[J]. International Journal of Computer Vision, 1999.

[2] 佟帅, 徐晓刚, 易成涛, 等. 基于视觉的三维重建技术综述[J]. 计算机应用研究, 2011, 28(7):2411-2417.

[3] YANG Lei, TIAN Bo. Perspective SFS 3-D Shape Reconstruction Algorithm with Hybrid Reflectance Model[J]. International Conference on Computer Science and Network Technology, 2011: 1764-1767.

[4] 杨志明,赵红东.从阴影恢复形状的径向基函数反射模型研究[J].中国图象图形学报,2017,22(11):1565-1573.

[5] 吴仑, 王涌天, 刘越. 一种鲁棒的基于光度立体视觉的表面重建方法[J]. 自动化学报, 2013, 39(8):1339-1348.

[6] XU Di, CAI Jianfei, ZHENG Jianmin, et al. Photometric stereo using mesh face based optimization[J]. Visual Communications and Image Processing,2016:27-30.

[7] Shuhei Hashhimoto, Daisuke Miyazaki, Shinsaku Hiura. et al. Uncalibrated photometric stereo constrained by intrinsic reflectance image and shape from silhouette[J]. 2019 16th International Conference on Machine Vision Application,2019:1-6.

[8] 付琳,洪海波,王晰,等. 基于逆向反射模型的非朗伯光度立体视觉[J]. 光学学报,2020,40(5):151-161.

[9] 陈明汉,任明俊,肖高博,等. 基于多尺度卷积网络的非朗伯光度立体视觉方法[J]. 中国科学:技术科学,2020,50(3):323-334.

[10] BAUMGART B G. Geometric modeling for computer vision[D]. Stanford University.1974.

[11] POTMESIL M. Generating octree models of 3D objects from their silhouettes in a sequence of images[J]. Computer Vision Graphics and Image Processing, 1987, 40(1):1-29.

[12] FORBES K, NICOLLS F, DE JAGER G, et al. Shape-from-silhouette with two mirrors and an uncalibrated camera[C]// European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2006: 165-178.

[13] HARO G. Shape from Silhouette Consensus[J]. Pattern Recognition, 2012, 45(9).

[14] KANATANI K I, CHOU T C. Shape from texture: General principle[J]. Artificial Intelligence, 1989, 38(1):1-48.

[15] BLAKE A, MARINOS C. Shape from texture: Estimation, isotropy and moments[J]. Artificial Intelligence, 1990, 45(3): 323-380.

[16] BLOSTEIN D, AHUJA N. Shape from texture: integrating texture-element extraction and surface estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(12): 1233-1251.

[17] 刘军, 后士浩, 张凯, 等. 基于单目视觉车辆姿态角估计和逆透视变换的车距测量[J]. 农业工程学报, 2018, 34(13):78-84.

[18] SAHAY R R, RAJAGOPALAN A N. Dealing With Parallax in Shape-From-Focus[J]. Image Processing, IEEE Transactions on, 2011, 20(2):558-569.

[19] MAHMOOD M T. Nonlinear Approach for Enhancement of Image Focus Volume in Shape From Focus[J]. Image Processing, IEEE Transactions on, 2012, 21(5):2866-2873.

[20] SNAVELY N, SEITZ S M, SZELISKI R. Modeling the World from Internet Photo Collections[J]. International Journal of Computer Vision, 2008,80(2):189-210.

[21] MOULON P, MONASSE P, MARLET R. Adaptive structure from motion with a contrario, model estimation[C]// Asian Conference on Computer Vision. Springer Berlin Heidelberg, 2012:257-270.

[22] WU C. Towards Linear-Time Incremental Structure from Motion[C]// 3DV-Conference, 2013 International Conference on. IEEE Computer Society, 2013.

[23] SCHONBERGER J, FRAHM J. Structure-from-Motion Revisited[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016: 4104-4113.

[24] Cui H, Shen S, Gao X, et al. Batched Incremental Structure-from-Motion[C]// International Conference on 3D Vision (3DV), Qingdao, 2017:205-214.

[25] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[J]. Advances in neural information processing systems 2014: 2366-2374.

[26] David Eigen, Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[J]. arXiv preprint,2014:1411.4734.

[27] CHEN Weifeng, FU Zhao, YANG Dawei, et al. Single-Image Depth Perception in the Wild[J]. arXiv:1604.03901,2016.

[28] LUO Yue, REN Jimmy, LIN Mude, et al. Single View Stereo Matching[J]. arXiv:1803.02612,2018.