项目 2：SFM算法实现

中国科学技术大学
安徽合肥 230027
zyzhong@mail.ustc.edu.cn

摘要

由运动到结构的三维重建技术（SFM）作为双目立体视觉重建的一种方法，在实际应用中发挥着越来越重要的作用。本文首先介绍SFM的基本原理、研究现状及主要研究方法，然后由某一场景的多视角图像，通过SFM算法重建三维点云数据，并在三维软件中重建三维模型，最后分析实验的优势与不足。

1. 研究现状及主要研究方法

1.1. 国内外研究现状

三维重建（3D Reconstruction）概念最早由Roberts于 1963 年中提出，目的是从输入图像中获取真实物体的三维实体模型。随着技术的发展，三维重建的输出已经从原始的实体模型发展到如今的三维场景模型。在此发展过程中产生了很多关于三维重建的应用,例如军事仿真、机械工业、农业[1]、影视业、游戏业、建筑行业、医疗行业、数字文化遗产的保护、无人驾驶[2]、地图导航[3]、地图构建[4]、场景监控[5]、3D打印、三维人脸重建[6]、三维人体重建、以及近年来发展迅速的虚拟现实与增强现实产业，三维重建技术已成为计算机视觉方面的一个研究热点。图 1.1 展示了三维重建的一些应用实例，这些应用利用了三维模型呈现信息的多样性，以便实现特殊应用的目的。

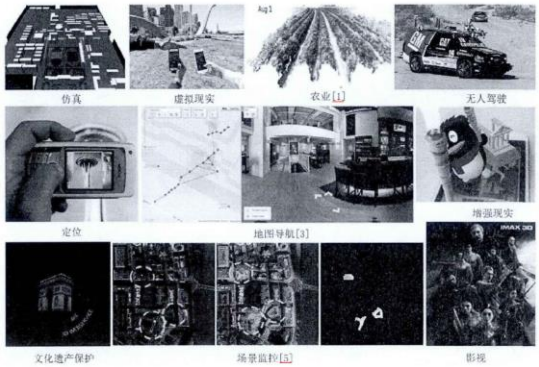

图 1.1 三维重建的应用实例[1,3,5]

Fig1.1 The application of 3D reconstruction

在对真实环境中的景物进行三维重建时，获取环境中物体三维模型的主要方法分为两种：主动式[7]和被动式[8-11]。

（1）主动式：主动式主要基于光学原理，对需要重建的物体进行光学扫描，从而得到目标的三维点云。随着相关技术的不断发展，到目前为止，该方式已逐渐形成了多种方法，包括结构光法[12]、激光扫描法[13]、阴影法等。

这种主动方式能获得更具体的物体信息，使重建后的物体细节更丰富，重建精度较高，是目前获取精确三维模型的较为通用的方法。尽管主动式有很多优势，但也存在着使其应用受到较大限制的不足。首先，该种方式需要使用扫描仪等设备，这使得其使用成本大大增加。此外，对于小规模的重建对象进行扫描是该技术可以实现的，但当需要对大规模对象或大范围景物进行重建时，无法对这么目标进行扫描等处理。因此，其应用的范围会受到设备等因素的限制，对重建物体的要求也较多。此外，对点云的后续分析不是十分简便。

（2）被动法：被动法通过所获取的重建对象图像来进行重构，利用图像所反映的数据，根据涉及的相关技术对其进行计算及转换等，最终得到重建目标三维信息。因此，这一过程也被称为基于图像的三维重建。该方法具有较高的灵活性，适用于各种复杂场景的三维建模。

在研究初期，主要针对单幅图像或一对图像，利用它们的信息完成重建工作。随着研究的不断深入以及计算机视觉领域中相关技术的发展，此类方法的研究重点逐渐向基于多幅图像的重建技术转移。目前，被动式的重建方法依然存在很多需要继续深入研究的问题。但相较其他建模方法，基于图像的三维重建仍有许多其他方法不可比拟的优势，总结如下：

1.可重建大规模场景：对重建规模及设备等没有十分严格或特殊的要求，不受这些因素的限制。所需设备即为一般的图像采集设备，如普通相机等。获取目标物体的二维信息后，即可通过对这些数据进行具体计算等来重建出目标物体。

2.重建模型具有较强的真实感：应用该方法重构出的三维模型能真实地体现真实景物的大小、外观以及凸显真实感的表面纹理，不需要再进行复杂的处理。

3.操作简单，易于实现：在建模的过程中，不需要耗费大量的精力和技巧，不需要人力的参与。

基于图像的三维重建技术自动化程度相对较高，适应性强，具备良好的应用前景。由二维图像来获得其对应的空间三维结构，在此过程中涉及到很多计算机视觉技术。为使三维重建技术更加成熟，各学者对这些关键技术进行了许多大量而深入的研究，也因此出现了许多的重建方法。其中，比较经典的有明暗法[14]，轮廓法[15]，运动法[16]等。本文研究工作主要关注的是从运动中恢复结构方法（Structure From Motion，SFM）。

SFM是由一系列包含着视觉运动信息（motion signals）的多幅二维图像序列（2D image sequences）估计三维结构（3D model）的技术，它属于计算机视觉及可视化的研究范围。

SFM通过图像集中的匹配点来估计三维静止场景中运动相机的内外参数和该场景相对于一个参考坐标系的结构关系，从而恢复 3D场景。

基于图像序列的三维重构是一个逐步获取关于被摄目标和摄像机信息的过程，第一步要通过对极关系按像对来依次建立各幅图像之间的联系，这一过程所获得的结果通常是数百或者数千个特征点在投影变换层次的三维坐标和摄像机的位置与姿态；第二步是进行摄像机的自动标定，摄像机姿态的失真通常会违背一个或者多个约束条件，通过摄像机的自动标定可以对投影重构进行变换，使所有约束条件都得到满足，从而获得图像序列的欧氏重构；第三步是对标定后的图像序列进行处理，得到一个被摄目标的密集的三维欧氏表面模型，同时可将图像纹理影射到模型表面，进而获得具有图像真实感的目标三维模型。

SFM算法的基本过程如图 1.2 所示。

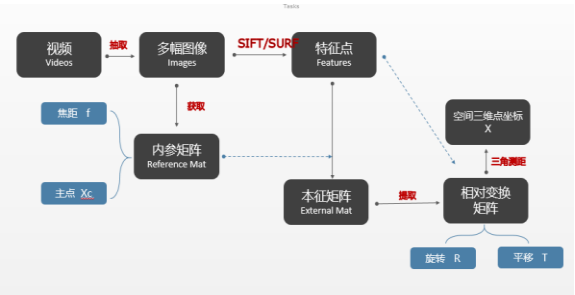

图 1.2 SFM算法的基本过程

Fig1.2 The pipeline of SFM algorithm

SFM的主要思想是首先利用摄像机采集景物的二维图像（未标定），通过对二维图像的匹配得到图像间的对应关系，再利用三维重建所涉及的关键技术进一步分析与计算，获取摄像机的内参与外参以及景物的空间三维数据。该方法不依赖于特定的假设条件，应用范围广。基于SFM的重建可分模块进行，即可将整个重建的过程分成若干部分。这种各部分可单独实现的模式便于控制与改进，且对环境等干扰因素具有鲁棒性。由此可知，该方法有一定的研究意义，且具有较高的实用价值。

1.2. 主要研究方法

就国内外目前研究现状来，基于SFM的三维重构主要包括图像匹配、摄像机标定和三维重建，重建过程为从两幅或多幅不同角度的图像中寻找匹配点，即x_1，x_2...（不同图像上的匹配点），并根据获得的匹配点二维坐标及相关几何约束关系得出对应的三维点坐标、摄像机内参及位置参数等。具体实现步骤如下：

（1）特征点检测与匹配，得到初始的特征点匹配结果；

（2）摄像机标定，以上步得到的匹配点坐标为基本数据，利用相关计算机视觉理论，进一步分析并计算出摄像机内、外参数；

（3）重建空间点三维信息，获得摄像机投影矩阵并与前面所得的可靠匹配点联立求解出空间点的三维坐标，得到三维点云。本文重点就其中的几个关键步骤展开研究。

1.2.1.特征点提取

特征点提取及匹配是SFM算法的基础，其目标是要从两幅图像中找到并正确匹配同一物体或场景的关键点。通常特征点的匹配包括特征点的描述和描述符向量间的匹配。就三维场景重建这一应用来说，效果较好且具有广泛的应用范围的特征点检测算法当属Harris角点提取算法[17]和SIFT特征[18]。

Harris[17]等人首先提出了角点的定义，该算子提取的特征点均匀合理，但是精度只能达到一个像素，不具备尺度不变性。

Harris角点提取算法的效果较好，该算子既可以保留大量重要的特征，又可以降低数据量。由于此算法的计算量相对较小，具有很好的稳定性，已经被广泛应用到各个相关领域。Harris角点能够较为准确的表示出物体的结构，但在图像局部，容易出现角点聚集的情况，该种情况下的角点在进行图像匹配时易导致大量误匹配的产生，并影响匹配速度与效果；此外，角点是二维图像灰度变化最明显的点或是边缘信息曲率极值的点，对一些干扰因素，如噪声等，具有较强的鲁棒性。但通常情况下，匹配此类角点时，一般采用灰度相关法。一般来说，这种方法可以得到相对准确的匹配点。但当应用的图像灰度非线性变换时，在这些干扰因素下所得匹配点中会包含一些误差较大甚至是错误的点对。因此，该算法在具有旋转或缩放等变换的图像匹配应用中的效果不是太理想。在实际应用中，因实际需求的逐渐增多，这些不足对其应用的限制也逐渐显现出来。

2004 年，Lowe等人正式提出了对于光照、尺度、旋转都有很强鲁棒性的SIFT算法 18]，该方法在图像二

维平面空间和高斯差分（DoG）尺度空间中同时检测局部极值作为特征点，以特征点为中心利用梯度直方图统计领域像素的梯度方向并采用 128 维的矢量描述图像。

基于尺度空间的SIFT算法首先在尺度空间内检测出关键点。然后，统计关键点一定邻域窗口内的梯度方向。根据所得信息数据，绘制直方图。最后，根据获得的直方图构造特征描述向量。这种方法在处理具有各类变换的图像时，具有较高的鲁棒性，应用效果较好，同时还具有可扩展性。SIFT算法具有一定的稳定性，其提取算子在亮度、尺度变化等干扰的情况下，仍然可以检测到大量准确的特征点。

本文使用光流法进行特征提取。光流是匹配来至一幅图像选择的点到另外一幅图像选择点的过程，假定这两个图像是一个视频序列的一部分并且它们彼此非常相近。大多数的光流方法比较一个小的区域，称为搜索窗口或者块，这些块围绕着图像A中的每一点和同样区域的图像B中的每一点。遵循计算机视觉中一个非常普通的规则，称为亮度恒定约束（brightness constancy constraint），图像中的这些小块从一个图像到另外一个图像不会有太大的变化，因此，他们的幅值差接近于 0。除了匹配块，更新的光流方法使用一些额外的方法来获得更好的结果。其中一个方法就是使用图像金字塔，它是图像越来越小的尺寸（大小）版本。另外一个方法是定义一个流场上的全局约束，假定这些点相互靠近，向同一方向一起运动。

1.2.2.摄像机标定

摄像机标定主要是确定摄像机的图像坐标系与物体空间中的三维坐标系之间的关系。

传统的摄像机标定方法有代表性的有直接线性变换法（DLT）、两步法、张正友标定法等。

Faugcras[22]等提出的自标定方法是利用绝对二次曲线的对极几何关系建立的Kruppa方程来标定相机的 5 个内参数。

1.2.3. 匹配点优化和基础矩阵的估计

基础矩阵它反映的是从一幅图像上的点x(x’)到另一幅图像上与之对应的对极线I’(I)的映射。求基础矩阵F，常用的算法是用RANSAC自动估计两幅图像之间的基础矩阵。Faugeras[21]证明了只要两幅图像间的基础矩阵已知，就可以实现射影重叠。

由摄像机对某目标进行多角度拍摄，获得的图像间存在着几何约束关系。在代数中，这些关系可由基础矩阵表示。基础矩阵可以由匹配点信息求出，并可用来恢复摄像机参数及重建目标的空间三维信息。由此可知，基础矩阵的精确求取是实现重建的纽带。因此，基础矩阵的鲁棒性参数估计是三维重建中非常重要的环节。到目前为止，其估计方法主要有线性方法、迭代方法及鲁棒方法。现阶段，比较流行的为鲁棒性较好的RANSAC算法。此外，RANSAC算法能在求解出基础矩阵的同时，可以剔除图像匹配过程所引入的部分误匹配，应用效果较好。

2. 本文研究方法

2.1. 获取相机的内参矩阵F

对输入图像利用其内部的编码信息获取相机的焦距，计算内参矩阵。

2.2. 图像预处理

在获取图像时，采用给定目录，逐个读取目录中的图像，对图像进行灰度化处理。

2.3. 特征点提取，对相邻图像两两计算匹配特征点

对输入图像进行角点检测。首先用两张图像作为初始化建出来一个初始的点云，之后不断添加后续的图像进入，并添加入点云。 那么具体添加那一张图像可以采用的方法是：看已有的图像中哪一个与已有点云中的点匹配最多就先选哪张。

2.4. 计算两图像之间的内参矩阵F

关于fundamental matrix的推导可以使用Multiview geometry 242 页 9.2.1 中的几何推导以及 9.2.2 中的算术推导，算术推导证明如果两张图像的拍摄是纯相机平移的话，fundamental matrix是计算失败的。计算内参矩阵可以使用 8 点算法。具体内容参看Multiview geometry 282 页 Algorithm 11.1，值得注意的是，在进行 8 点算法之前，需要使用RANSAC算法对特征点进行提纯（RANSAC算法中的模型使用计算基础矩阵的 8 点算法，参见Multiview geometry 121 页 Algorithm 4.5），同时在输入计算基础矩阵之前，需要对所有的特征点进行normalize，这里是必须要做的，原因参见Multiview geometry 108 页，具体做法使用 109 页 Algorithm 4.2.

2.5. 计算相机的基础矩阵E

使用公式 $E=K^{'T}FK$，其中K与$K^{'T}$分别对应于两个视角的内参矩阵。

基础矩阵（用F表示）和本征矩阵（用E表示）。在得到匹配筛选过的特征点后，就能够计算出图像间的本征矩阵了。使用OpenCV中的findEssentialMat()方法可以直接实现。之后对求得本征矩阵进行SVD奇异值分解，得到旋转部分和平移部分。进行分解的E，得到R1，R2，T1，T2。

2.6. 计算[R|t]矩阵

利用基础矩阵E计算两个相机之间的外参，即[R|t]矩阵。

这里参见Multiview geometry 258 页 9.6.2。但是由于符号的关系，会输出出来 4 中可能的[R|t]矩阵，这个时候我们将所有 2D的点利用这四种[R|t]映射到 3D空间中去，看哪一种[R|t]对应的 3D点的z深度方向全部是正向的。因为准确的[R|t]场景点都在相机朝向的正前方。

2.7. 利用三角测量原理进行三维构建

结构光测量中为了获取物体的三维信息，一般都会使用三角测量原理；其基本思想是利用结构光照明中的几何信息帮助提供景物中的几何信息，根据相机，结构光，物体之间的几何关系，来确定物体的三维信息。

2.8. 优化匹配组

目前我们已经知道两两图像之间的[R|t],比如 1、2之间的，2、3之间的，如何将 1、2、3放入同一个参考坐标系，需要使用BA——数学上是最小化投影误差，让所有得出的[R|t]与 3D点估计输出的 2D点与实际观测点位置最小化。

对特征点进行优化，基本去除错误匹配，进行重新构建。

3. 实验结果

本节采用算法检测提取实验图像中的特征，并选取其中的两张图像作为样例，展示基于SIFT的匹配算法效果。

图 3.1 不同拍摄角度的两张图像

Fig3.1 Two images at different shooting angles

图 3.2 图像初始匹配结果

Fig 3.2 Initial image matching results

图 3.2 是经算法匹配后的实验结果。图中的线段代表将两幅图像特征点进行匹配的连接线。线段的两个端点是一对匹配点，其中与多数平行直线相交的直线为误匹配。由实验结果可知，基于SIFT的匹配算法能较好地对提取出的特征点进行匹配，但结果中仍存在着一些误匹配。这些误匹配会影响后面工作的进行，降低三维重建的精度，需要剔除。

```
-0.0471062 -0.966252 5.12672
-0.372026 -0.767094 5.33449
-0.0376253 -0.775412 5.13765
-0.210726 -0.773979 5.24775
1.17513 -0.0223665 2.40726
1.10065 -0.0385795 2.29724
-0.514245 -0.77695 5.39929
-0.40048 -1.03467 5.31794
0.482223 -0.152435 3.65204
-0.0933688 -1.03751 5.15158
-0.667826 -1.02044 5.42544
0.941688 -0.209559 3.56413
-0.36713 -0.275124 5.452
0.00762081 -0.148307 4.05867
-1.15531 -0.446296 5.63326
1.12254 0.173957 3.30921
0.44004 -0.202008 3.72276
-0.679885 -0.972157 5.43519
-0.661514 -0.266639 5.62184
1.22127 -0.174387 2.51713
1.26088 -0.0297514 2.59852
0.656506 -1.1479 4.53775
-0.0196333 -0.375732 4.36906
0.0386273 -0.556497 5.10664
-0.828529 -1.37801 5.43092
0.177163 -1.08788 4.96044
-0.548763 -0.424664 5.47797
0.462977 -0.114967 3.69795
-0.310124 -0.403901 5.34275
-1.06291 -0.44529 5.63746
-0.261523 -0.657734 5.29528
1.07955 0.161577 3.42355
-0.263114 -0.896882 5.26516
0.864252 -0.884097 4.26329
0.322425 -1.09353 4.8502
```

图 3.3 重建三维点云结果（部分）

Fig3.3 The results of 3D point cloud reconstruction

利用三维点云数据在三维软件MeshLab中重建三维模型，结果如图 3.4 所示。

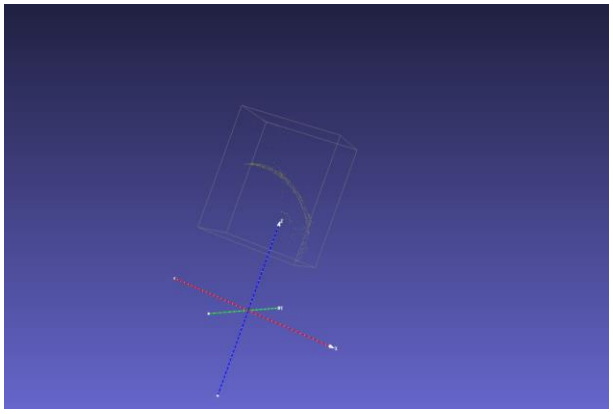

图 3.4 MeshLab重建三维模型

Fig3.4 MeshLab reconstruction 3D model

4. 总结

优点：SFM算法对图像的要求非常低，可以采用视频图像序列进行三维重建。可以使用图像序列在重建过程中实现摄像机的自标定，省去了预先对摄像机进行标定的步骤，由于各种特征点提取和匹配技术的进步，运动法的鲁棒性也极强。运动法的另一个巨大的优势是可以对大规模场景进行重建，输入图像数量也可以达到百万级。

缺点：运算量比较大，同时由于重建效果依赖特征点的密集程度，对特征点较少的弱纹理场景的重建效果比较一般。

参考文献

[1] Dong, Jing, et al. "4d crop monitoring: Spatio-temporal reconstruction for agriculture." Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017.

[2] Song, Shiyu, and Manmohan Chandraker. "Robust scale estimation in real-time monocular SFM for autonomous driving." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[3] Colbert, Mark, et al. "Building indoor multi-panorama experiences at scale." ACM Siggraph 2012 Talks. ACM, 2012.

[4] 刘浩敏, 章国锋, 鲍虎军. "基于单目视觉的同时定位与地图构建方法综述." 计算机辅助设计与图形学学报 28.6 (2016): 855-868.

[5] Lin, Baowei, et al. "Image Based Detection of 3D Scene Change." IEEJ Transactions on Electronics, Information and Systems 133.1 (2013): 103-110.

[6] 王琨, 郑南宁. "基于 SFM 算法的三维人脸模型重建." 计算机学报 28.6 (2005): 1048-1053.

[7] Hartley, Richard I. "Self-calibration of stationary cameras." International journal of computer vision 22.1 (1997): 5-23.

[8] Hauswiesner, Stefan, Matthias Straka, and Gerhard Reitmayr. "Temporal coherence in image-based visual hull rendering." IEEE transactions on visualization and computer graphics 19.10 (2013): 1758-1767.

[9] Li, Jianguo, et al. "Bundled depth-map merging for multi-view stereo." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[10] Muller, Karsten, Philipp Merkle, and Thomas Wiegand. "3-D video representation using depth maps." Proceedings of the IEEE 99.4 (2011): 643-656.

[11] Taveira, Giancarlo, and Leandro AF Fernandes. "Automatic alignment and reconstruction of facial depth images." Pattern Recognition Letters 50 (2014): 82-90.

[12] Rocchini, C. M. P. P. C., et al. "A low cost 3D scanner based on structured light." Computer Graphics Forum. Vol. 20. No. 3. Blackwell Publishers Ltd, 2001.

[13] Huber, Daniel F., and Martial Hebert. "3d modeling using a statistical sensor model and stochastic search." Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Vol. 1. IEEE, 2003.

[14] Ecker, Ady, and Allan D. Jepson. "Polynomial shape from shading." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[15] Haro, Gloria, and Montse Pardàs. "Shape from incomplete silhouettes based on the reprojection error." Image and Vision Computing 28.9 (2010): 1354-1368.
[16] Szeliski, Richard. Computer vision: algorithms and applications. Springer Science & Business Media, 2010.
[17] Harris C,Stephens M. A combined conor and edge Detector[C]. Processing of the Fourth Alvey Vision Conference. 1998:147-151.
[18] Lowe. Object recognition from local scale-invarant features[C]. Processing of the 7th IEEE International Conference on Computer Vision. Greece, 1999:1150-1157.
[19] Hilton A,Illingworth J. Geometric fusion for a hand-held 3D sensor[J]. Machine Vision & Applications, 2000, 12(1):44-51.
[20] Cui Y. 3D shape scanning with a time-of-flight camera[J]. Computer Vision and Pattern Recognition(CVPR), 2010, 23(3):1173-1180.
[21] Arce G R. Nonlinear signal processing: A statistical approach[M]. Hoboken: John Wiley & Sons, Inc, 2005:80-138.
[22] Nixon M S, Aguado A S. Feature extraction and image processing[J]. Journal of Medical Ethics, 2008, 26(1):78.
[23] Ezra E, Sharir M, Efrat A. On the Performance of the ICP Algorithm[J]. Computational Geometry, 2008, 41(1-2):77-93.
[24] Sharp G C,Lee S W,Wehe D K. ICP registration using invariant features[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(1):90-102.