基于多尺度几何一致性的多视图立体匹配

徐梓轩
21821170
计算机科学与技术学院
xzx_marco@qq.com

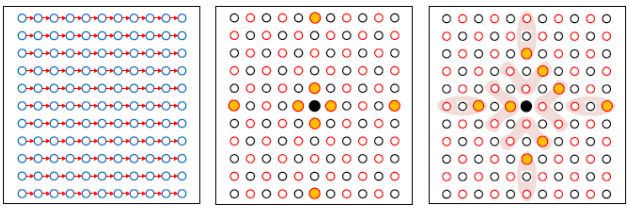
摘要（Abstract）

本文介绍了一个新颖的多视图立体匹配算法，它使用多尺度几何一致性对深度图的估计进行引导。该算法的贡献主要分为两个部分，第一部分为适应性棋盘采样及多假设联合视图选择，能够充分利用结构性区域信息，在传播过程中进行更好的采样，以实现鲁棒性更高的视图选择；第二部分即为多尺度几何一致性引导，是对第一部分的进一步完善，使得算法能够实现对弱纹理区域的深度估计，利用了小尺度下能够更好地估计弱纹理区域深度的特性，使用小尺度估计的深度图引导大尺度下深度图的估计。大量的实验证明，该方法的表现超过了现有方法，不管在弱纹理区域还是细节程度高的区域，都能够估计较为准确的深度。

1. 引言

近年来，三维重建在文物数字化，动画电影，游戏等行业中的应用越来越广泛，它能从二维图像集中恢复出对应场景的三维结构。传统的三维重建流水线包括稀疏点云重建，密集点云重建，面片重建及纹理映射，其中，密集点云重建主要通过多视图立体匹配（Multi-view Stereo, MVS）进行实现。MVS是一个传统的计算机视觉话题，其目标是从已知对应相机姿态的图像集中恢复场景的密集三维点云。近几年，许多工作致力于提升MVS结果的质量并取得了不错的成果，但在数据规模很大，图像中存在弱纹理，遮挡，重复结构，以及高光反射等情况下，如何进行准确的，高效的多视图立体匹配仍然是一个十分具有挑战性的问题。

MVS的方法大致可以分为四类，分别基于体素，面片演化，图像块和深度图。其中，基于深度图的方法是目前的研究热点，属于这一类的PatchMatch立体匹配方法[1,2,3,4]已经展现了其重建密集点云的潜力，该方法



(a) (b) (c)

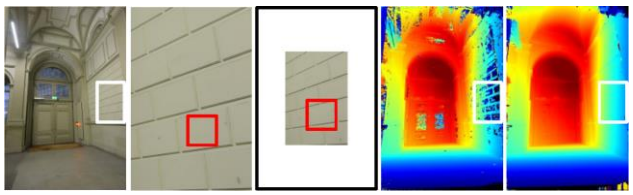
图 1 不同的传播模式 (a) 顺序传播； (b)对称性棋盘传播；(c) 适应性棋盘传播

一般由四个步骤组成，分别为随机初始化，传播，视图选择以及优化。其中，传播和视图选择是PatchMatch方法中的关键步骤，前者影响该方法的效率，而后者则决定该方法的准确性。

在传播过程中，每个像素需要根据其邻域内其他特定像素进行调整，目前大体上存在两种传播模式，一种是顺序传播[1,2,4]，另一种是扩散传播[3]。顺序传播中，每次传播沿着特定方向进行，如图 1(a)所示；扩散传播则如图 1(b)所示，黑色实心圆圈所表示的像素根据黄色实心像素进行更新。就并行性而言，后者能一次性更新一半的像素（分别用红色和黑色描边表示），而前者一次只能更新一行或一列像素，显然后者效率更高。然而，在视图选择的实现上，后者[3]却不像前者[2,4]那样精细，导致后者的视图选择是有偏的，在某些情况下整体表现不如顺序传播的方法。

在使用棋盘（即扩散）传播的前提下，为了获得更好的重建质量，需要设计一个更加鲁棒的视图选择方法，本文将要介绍的MVS方法[5]便尝试解决这个问题。

该方法的贡献主要包含两部分，第一部分是一个基础的MVS方法，其中使用了适应性棋盘采样和多假设联合视图选择（原文简称为ACMH），这样的策略能够利用图像中的结构性区域信息；第二部分则试图解决弱纹理的问题，如图 2(b)所示，红色方框表示图像块窗口，其中包含了弱纹理区域，弱纹理使得图像间像素点的匹配变得困难，从而影响视图选择的准确性，进而影响深度图的质量（如图 2(d)所示）。但当我们对图像进行下



(a) (b) (c) (d) (e)

图 2 不同尺度下的纹理复杂度 (a) 原图像； (b) 图(a)中白色方框内容的放大图； (c) 图(b)的下采样图； (d) 初始尺度下获得的深度图； (e) 多尺度模式下获得的深度图

采样后，保持窗口大小不变，从图 2(c)可以看到，窗口中包含了部分纹理信息，可见，纹理复杂度是一个相对的概念。因此，我们首先可以估计低分辨率图像的深度图，以减轻弱纹理带来的问题，接着用低分辨率的深度图来引导高分辨率深度图的计算。第二部分称为基于几何一致性引导的多尺度图像块匹配（原文简称为ACMM）。

结合这两部分，使得该方法在保留扩散传播模式的前提下，进行更加准确的视图选择，进而获得质量更高的深度图。

2. 算法框架

给定一个图像集$\mathcal{I}=\{I_i|i=1\dots N\}$，对应的相机参数集为$\mathcal{P}=\{P_i|i=1\dots N\}$，算法的目标是估计每张图像的深度图$\mathcal{D}=\{D_i|i=1\dots N\}$，并将它们融合成一个三维点云。每次按顺序从$\mathcal{I}$中取一张图像作为参考图像$I_{ref}$，以剩下的图像作为源图像$I_{src}$进行深度估计。

该算法[5]的框架如图 3 所示，整个框架即为ACMM，首先对所有图像进行$k-1$次下采样，记图像I_i的第l个尺度及其对应的相机参数为I_i^l和$P_i^l, l=0\dots k-1$，I_i^{k-1}则为原图像。使用多尺度的目的是将小尺度下对弱纹理区域的深度估计传播到大尺度中，并引导大尺度下的深度估计，同时保留细节。

接着，从尺度最小的图像开始，先进行随机初始化，再进行ACMH，此时使用光度一致性对I_{ref}和I_{src}之间像素的相似性进行衡量，这一部分将在第 3 节中详细介绍。为了增强所有深度图之间的一致性，还需要进行第二次ACMH，但此时使用几何一致性进行衡量。对结果进行上采样和细节恢复，得到下一尺度的深度图，再进行基于几何一致性的ACMH，这样能够在当前尺度下保留并优化上一尺度中对于弱纹理区域较为可靠的估计，这一部分将在第 4 节中详细介绍。重复这些步骤，直到得到原始尺度下的深度图。

3. 结构性区域信息

结构性区域信息是适应性棋盘采样和多假设联合视图选择（即ACMH）中主要使用的性质，即图像上某一相对较大的区域中的像素可以使用同一个三维平面进行建模。利用该信息使得算法能够对更好的候选假设进行采样以便传播，并选择可信度更高的视图。接下来介绍[5]中基础MVS方法（即ACMH）的四个具体步骤。

3.1. 随机初始化

首先为参考图像I_{ref}中的每一个像素随机生成一个假设，该假设包含了深度和法向量，使用这两个信息以及I_{ref}对应的相机参数，可以在三维空间中唯一确定一个平面。接着，对于每个假设，使用剩下$N-1$张源图像逐一计算匹配代价，并从中取出最小的K个匹配代价进行聚合，成为该假设的初始多视图匹配代价，为接下来的传播过程做准备。

3.2. 适应性棋盘采样

首先，适应性棋盘采样与[3]（如图 1(b)所示）一样，将I_{ref}中的像素划分为红色和黑色描边两部分，黑色像素仅利用红色像素的信息，因此可以一次性更新所有的黑色像素，反之亦然。在[3]中，每个像素在固定的 8 个位置上进行采样，而适应性方法将这样的采样模式进行扩展。如图 1(c)中的红色底纹所示，每个像素有 8 个采样区域，包括 4 个V型区域和 4 个长条区域，在每个区域中，根据像素的多视图匹配代价来选取一个最好的样本（图中黄色实心像素），样本的假设记为$h_i, i=1,2\dots,8$。这种采样策略能够更好的利用结构性区域信息，使得一个更好的三维平面能够包含尽量多的像素。

3.3. 多假设联合视图选择

为了进一步优化I_{ref}中每个像素的多视图匹配代价，需要利用上述 8 个样本来选择新的视图。对于像素p，利用其样本对应的假设来构造一个代价矩阵

$$M=\begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,N-1} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{8,1} & m_{8,2} & \cdots & m_{8,N-1} \end{bmatrix} \tag{1}$$

其中$m_{i,j}$表示使用第j个视图I_j计算得到的第i个假设h_i的匹配代价，匹配代价使用归一化互相关的双边加权调

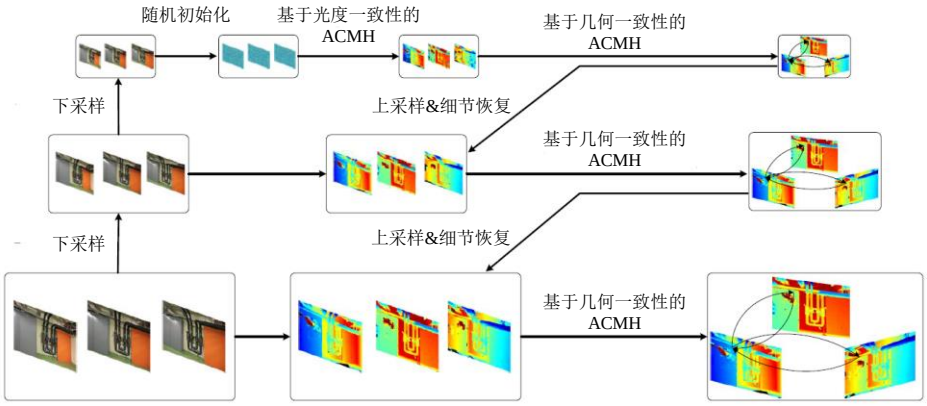


图 3 本文介绍方法的框架

整[4]来计算，该计算方法描述了参考图像块和源图像块之间的光度一致性。

接着根据M选取用于聚合的视图，作者通过观察认为，对于一个较差的视图来说，它对应的 8 个匹配代价通常都很大，反之，一个好的视图有着较小的匹配代价，且会随着迭代的进行不断降低。因此，使用投票机制进行视图选择，该机制使用了两个阈值τ_1和$\tau(t)$，其中，$\tau(t)$的定义为

$$\tau(t) = \tau_0 \cdot e^{-\frac{t^2}{\alpha}} \tag{2}$$

其中t表示第t次迭代，τ_0表示初始的匹配代价阈值，α是一个常数；另一个阈值τ_1是常数，且有$\tau_1 > \tau(t)$。基于上述观察，若需要选择视图I_j，则其应满足以下两个条件：(a) 存在多于n_1个匹配代价满足$m_{i,j} < \tau(t)$，这些匹配代价的集合记为$S_{good}(j)$；(b) 少于n_2个匹配代价满足$m_{i,j} > \tau_1$。此时，I_j将被加入到当前迭代t对应的视图集合S_t中。

S_t中可能包含了一些效果相对较差的视图，因此需要为集合中的每个视图赋予权重。首先，定义匹配代价$m_{i,j}$的可信度为

$$C(m_{i,j}) = e^{-\frac{m_{i,j}^2}{2\beta^2}} \tag{3}$$

其中β为常数。于是，视图I_j的权重定义为

$$\omega(I_j) = \frac{1}{|S_{good}(j)|} \sum_{m_{i,j} \in S_{good}(j)} C(m_{i,j}), I_j \in S_t \tag{4}$$

假设第$t-1$次迭代中权重最高的视图v_{t-1}对当前迭代t仍然起着一定的作用，则将式(4)改为

$$\omega'(I_j) = \begin{cases} (\mathbb{I}(I_j = v_{t-1}) + 1) \cdot \omega(I_j) & \text{if } I_j \in S_t \\ 0.2 \cdot \mathbb{I}(I_j = v_{t-1}) & \text{else} \end{cases} \tag{5}$$

其中$\mathbb{I}(\cdot)$是指示函数，即有$\mathbb{I}(\text{true}) = 1$且$\mathbb{I}(\text{false}) = 0$。

利用式(5)，可以进一步计算像素p对于每个样本假设h_i的多视图光度一致性匹配代价，其定义为

$$m_{photo}(p, h_i) = \frac{\sum_{j=1}^{N-1} \omega'(I_j) \cdot m_{i,j}}{\sum_{j=1}^{N-1} \omega'(I_j)} \tag{6}$$

最终，从 8 个样本中选出匹配代价最小的一个，用它的假设来更新像素p的假设。

3.4. 优化

每一次迭代后，需要进行一次优化以提高解空间的多样性。对于像素p当前的假设，即深度和法向量，存在三种情况：其中一个接近最优解，两个都不接近以及两个都接近[4]。因此，可以生成两个新的假设，其中一个随机生成，而另一个通过对当前假设进行扰动得到。对这些深度和法向量进行组合，最终可以得到 9 个假设，并从中选择多视图匹配代价最小的一个作为像素p的最终假设。

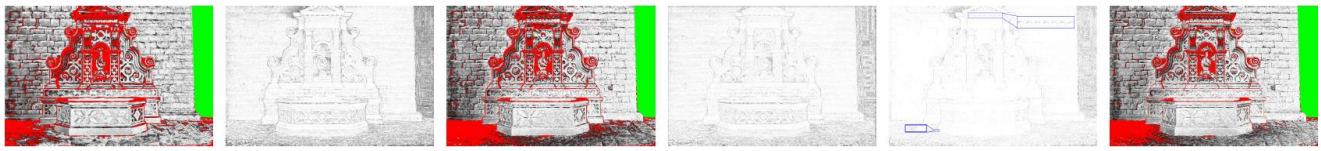
4. 多尺度几何一致性

当尺度变大时，使用光度一致性难以对小尺度下估计的深度进行优化，因此需要引入几何一致性。进行基于几何一致性的ACMH后，为了减少几何一致性及上采样导致的细节丢失，还需要进行细节恢复。

4.1. 几何一致性

使用前向后向重投影误差[4,6]来衡量几何一致性，设图像I_i中某像素p的深度为$D_i(p)$，且图像对应的相机参数为$P_i = [M_i | p_{i,4}]$，将像素p投影回三维空间中，得到的三维点坐标为

$$X_i(p) = M_i^{-1} \cdot (D_i(p) \cdot p - p_{i,4}) \tag{7}$$



(a) (b) (c) (d) (e) (f)

图 4 绝对误差图以及光度一致性代价图 (a) 使用不带细节恢复的ACMM时的绝对误差图，对应深度图由倒数第二个尺度的深度图进行上采样获得；(b) 图(a)的光度一致性代价图；(c) 使用基于光度一致性的ACMH时的绝对误差图；(d) 图(c)的光度一致性代价图；(e) 图(b)和图(d)之间的差；(f) 使用带细节恢复步骤的ACMM时的绝对误差图。绝对误差图中，红色像素表示误差大于 2cm，绿色像素表示缺失真实数据，误差在 0 到 2cm之间的像素的灰度值为[255,0]。

则参考图像I_{ref}中某像素p的第i个样本的假设与源图像I_j之间的重投影误差为

$$\Delta e_{i,j} = \min\left(\left\|P_{ref}\cdot X_j\left(P_j\cdot X_{ref}(p)\right)-p\right\|,\delta\right) \quad (8)$$

其中δ是一个阈值，防止出现遮挡时重投影误差过大。将式(8)整合到式(6)中，得到多视图几何一致性匹配代价

$$m_{geo}(p,h_i)=\frac{\sum_{j=1}^{N-1}\omega'(I_j)\cdot\left(m_{i,j}+\lambda\cdot\Delta e_{i,j}\right)}{\sum_{j=1}^{N-1}\omega'(I_j)} \quad (9)$$

其中λ用于权衡两项。

在第l个尺度下，使用联合双边上采样将上一个尺度的深度估计传播到下一个尺度中，并作为当前尺度ACMH的输入，但此时使用式(9)而不是式(6)来更新像素的假设。这样可以限制当前尺度下像素假设（尤其是弱纹理区域）的解空间，使得小尺度下对弱纹理区域可靠的深度估计能够传播到大尺度中。

4.2. 细节恢复

使用多尺度几何一致性引导深度图估计会导致深度图中的细节变得模糊，另外，上采样也会引入错误的深度值。为了解决这一问题，作者观察到，仅使用光度一致性时能够获得更好的细节（如图 4(c)所示），因此，可以考虑使用光度一致性来探测深度图中对细节的错误估计并进行修正。

如图 4(a)所示，细节丢失一般发生在狭长结构及边界上，为了追踪这些细节，根据图 4(e)，相邻两个尺度光度一致性代价图之间的差值能够突出细节部分的误差，而那些弱纹理区域的差值几乎为 0。因此，对上一尺度$l-1$的深度图进行上采样后，计算其光度一致性代价图，记为C_{init}^{l}，然后在当前尺度l使用彩色图像执行基于光度一致性的ACMH，得到另一张光度一致性代价图C_{photo}^{l}，则对于像素p，若满足

$$C_{init}^{l}(p)-C_{photo}^{l}(p)>\xi \quad (10)$$

其中ξ用于控制宽容度，则认为其深度值错误，并根据两者的差进行修正。使用细节恢复后的结果如图 4(f)所示。

当如图 3 所示的整个算法执行完毕后，需要融合所有估计的深度图，以得到一个密集的三维点云。

5. 实验

文章作者主要进行了两部分实验，第一部分验证深度图估计的准确性，在ETH3D基准测试中，与当前的主流方法COLMAP[4]进行了比较，ACMH几乎在所有数据集中都得到了次好的成绩，而ACMM更是在所有数据集中表现最好的。第二部分验证三维点云的准确性和完整性，与PMVS[7]，LTVRE[8],COLMAP等方法进行比较。总体而言，虽然大多数时候ACMM的准确性不是最好的，但差别不大。至于完整性，ACMM的表现十分出色，且能与其他方法拉开差距。因此，ACMM的F1 分数在所有数据集中的平均表现是最好的。

另外，在运行时间上，ACMH比COLMAP快 5-6 倍，尽管ACMM使用了多尺度模式，但其运行时间只是ACMH的两倍左右，所以仍然比COLMAP要快。

6. 结论

本文介绍了一个新颖的多视图立体匹配算法，根据大量的实验可知，该方法不管在质量上还是在时间上的表现都十分出色。质量上的提升得益于鲁棒性较高的多假设联合视图选择方法，以及创新性地使用多尺度几何一致性引导，它使得算法能够在弱纹理区域较为准确地估计深度值。时间上的提升则得益于并行性更高的适应性棋盘传播方法。这些提升使得算法在获得质量较高的三维点云的同时，还能够保持相对较高的效率，展现了良好的应用前景。

References

[1] Christian Bailer, Manuel Finckh, and Hendrik P. A. Lensch. Scale robust multi view stereo. In Proceedings of the European Conference on Computer Vision, pages 398–411, 2012.

[2] E. Zheng, E. Dunn, V. Jojic, and J. M. Frahm. Patchmatch based joint view selection and depthmap estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1510–1517, 2014.

[3] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, pages 873–881, 2015.

[4] Johannes L. Schonberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision, pages 501–518, 2016.

[5] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pages 5483-5492, 2019.

[6] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Recovering consistent video depth maps via bundle optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008.

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8):1362–1376, 2010.

[8] Andreas Kuhn, Heiko Hirschmuller, Daniel Scharstein, and Helmut Mayer. A tv prior for high-quality scalable multiview stereo reconstruction. International Journal of Computer Vision, 124(1):2–17, 2017.