Abstract

视觉SLAM技术在过去的十年取得了飞速的发展，并在计算机视觉和机器人领域有着广泛的应用。但是，绝大多数的视觉SLAM算法都建立在静态环境的假设下。当环境因为物体运动而发生改变时，系统的定位精度将会受到很大影响。在这篇技术报告中，我们对现有针对运动场景的SLAM和SfM技术进行了调研。一部分算法将运动区域的输入数据作为外点（outlier）剔除，来维护静态世界的基本假设；另一部分算法则同时解决静态环境和动态物体的结构和姿态估计。在序言部分，我们总结了动态环境中运动估计违背静态世界假设的基本约束，以及这些基本约束在动态环境下的基本延伸。同时，我们整理归纳了针对动态场景的代表性SLAM算法。我们将文献综述归纳为两类：一类算法以传感器为考虑的重点,基于经典的SfM框架来解决这一问题；另一类算法以地图为考虑的重点，更侧重于维护合适的稠密地图来解决这一问题。

动态场景下的SLAM相关技术总结报告

作者 Author *

August 26, 2019

1 序言

1.1 非刚体和多刚体运动下的SfM技术

为了处理场景中的动态物体，如之前所述将场景中不同运动物体进行分割，并对这些不同运动的物体分别进行三维重建是一个比较直接的方案。但考虑到所有物体的运动和三维信息同时都反应到了视频序列中，理论上这些物体的运动和三维信息可以同时进行求解[1]。在给定特征对应或者像素对应关系的基础上，基于矩阵分解的方式可以从表示了图像序列的特征矩阵中同时求解出动态物体的分割、恢复出各自物体的运动信息以及场景和物体的三维信息。这些方法根据场景三维结构在相机运动的模型下生成图像序列的过程，推导出最终的特征矩阵的特性。根据不同物体运动不同在矩阵中反应出的不同性质，对矩阵进行重新组织，并可以根据图像序列的生成模型将每部分的矩阵分解乘包含了相机运动的矩阵乘以三维信息的形式，利用矩阵中提供的约束同时完成运动物体分割、运动求解以及三维信息恢复。

*作者介绍 Brief introduction

1.2 视频序列帧中特征变化的子空间约束

与一般对特征的处理相同，假如我们可以跟踪到视频序列中的一系列特征，比如使用光流等方法[2]，如图1所示。为了便于推导，我们在这里假设这些特征在1至f帧之间均连续观测到，噪声和缺失的情况会另外探讨。我们将观测到的特征点记做$x_{ij} = (u_{ij}, v_{ij})$，其中下标中$i$代表第$i$帧，$j$代表第$j$个特征点。由于这些特征点是由相机在三维空间中运动生成的，所以这一系列特征点应当连续变化并与三维场景和相机运动对应。我们将这些特征点的坐标根据编号和时间序列排布到一个矩阵中，如式(1)所示，纵坐标方向上按照帧的时间顺序排列，横坐标按照特征点的编号排列。

$$W = \begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \vdots & \vdots \\ u_{f1} & \cdots & u_{fp} \\ v_{11} & \cdots & v_{1p} \\ \vdots & \vdots & \vdots \\ v_{f1} & \cdots & v_{fp} \end{pmatrix} \tag{1}$$

矩阵分解方法认为这个矩阵是由相机的运动和三维结果矩阵合成而成，可以通过矩阵分解恢复出原始的信息。

通过矩阵分解的方式联合求解摄像机运动和三维信息，是SfM中一个重要的方法。这种方法具有优雅的数学描述，充分的考虑到了特征之间在空间和时间上的关联约束。这种方法最早由Tomasi和Kanade[1]根据秩理论在1992年提出。他们的理论指出，在一个面向静态场景的较短的序列中，包含了在整个帧序列上所有跟踪到的特征点的观测矩阵（measurement matrix），它的秩最多是4。特别的对欧式坐标系下的垂直投影来说秩最多是3[3, 1, 4, 5]。这种秩下的约束表明了这些特征点在时间变化是有明显关联的。

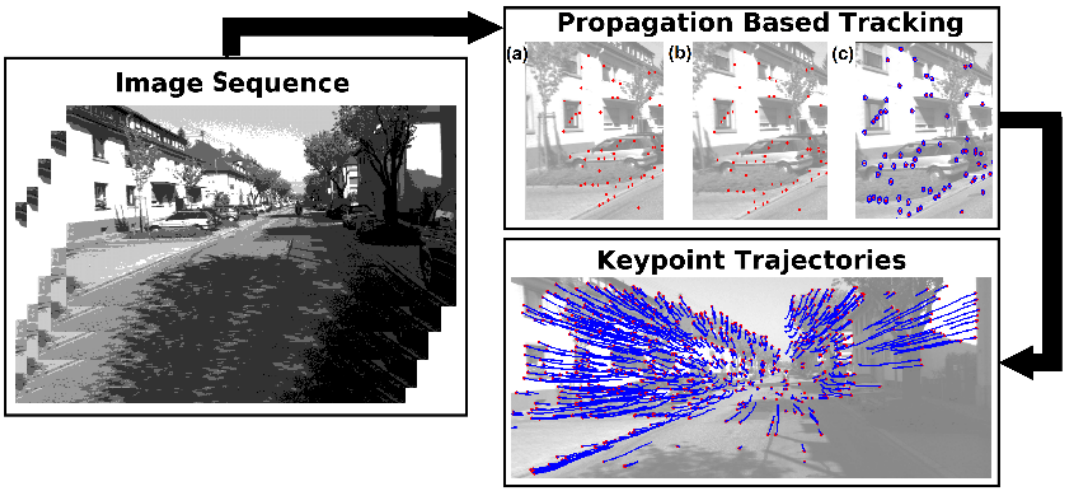


Figure 1: 跟踪得到的特征点序列轨迹[2]，观测矩阵是将图中连续出现的特征点坐标放到矩阵中，能够表示特征点在空间和时域上的关系。

1.3 观测矩阵与相机运动及三维结构的关系

观测矩阵是由相机的运动和三维点共通生成的，本节主要讲观测矩阵与这两个信息之间有什么样的关系。 由于垂直投影垂直投影形式较为简单，故我们先从垂直投影的情形来做说明[1]。垂直投影的形式如式(2)所示，是三维点的x, y分量的直接映射。

$$x_j = \begin{pmatrix} R_{11} & R_{12} & R_{13} & t_1 \\ R_{21} & R_{22} & R_{23} & t_2 \end{pmatrix} \begin{pmatrix} X_j \\ Y_j \\ Z_j \\ 1 \end{pmatrix} \tag{2}$$

其中R_{ij}是旋转矩阵R的第(i, j)个元素，t_i是平移分量的元素，上述矩阵中仅包含了旋转矩阵的前两行，第三行可以用这两行的叉乘求得。$[X_j, Y_j, Z_j, 1]$是第j个三维点齐次坐标，我们把第j个三维点的齐次坐标写作X_j，则对第n帧中的p个点来说，三维点的齐次坐标集合可以写作$[X_1, \cdots, X_p] \in \mathbb{R}^{4\times p}$。设$R_x^j$和$R_y^j$分别代表了第j帧姿态的旋转矩阵的第一行和第二行，$t_x^j$和$t_y^j$代表了第j帧姿态的平移分量。我们对每个相机对整个三维点集用式(2)进行投影，将得到的二维点堆叠起

来就可以得到式(1)中的观测矩阵，如式(3)所示。

$$\begin{pmatrix} u_{11} & \cdots & u_{1p} \\ \vdots & \vdots & \vdots \\ u_{f1} & \cdots & u_{fp} \\ v_{11} & \cdots & v_{1p} \\ \vdots & \vdots & \vdots \\ v_{f1} & \cdots & v_{fp} \end{pmatrix} = \begin{pmatrix} R_x^{1T} & t_x^1 \\ \vdots & \vdots \\ R_x^{fT} & t_x^f \\ R_y^{1T} & t_y^1 \\ \vdots & \vdots \\ R_y^{fT} & t_y^f \end{pmatrix} \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix} \tag{3}$$

一个更复杂的情况是在仿射相机（affine camera）的模型下进行推导[6]。在这个假设下，相机的投影模型可以简化成如式(4)所示的形式，旋转矩阵中最后一行为0。

$$x_j = \pi\left(\begin{pmatrix} f & 0 & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} & R_{13} & t_1 \\ R_{21} & R_{22} & R_{23} & t_2 \\ 0 & 0 & 0 & d_0 \end{pmatrix} \begin{pmatrix} X_j \\ Y_j \\ Z_j \\ 1 \end{pmatrix} \right) \tag{4}$$

其中f是相机的焦距，(c_1, c_2)是图像中心，$d_0 \in \mathbb{R}$ 是一个常量，R_{ij}是旋转矩阵R的第(i,j)个元素，t_i是平移分量的元素，$\pi(\cdot)$是讲齐次量变化为非齐次量的过程，即$\pi([X,Y,Z] = [X/Z, Y/Z])$。 我们定义如下的观测矩阵，其中每个位置为当前帧的二维位置减去第一帧的位置$x_{ij} - x_{1j}$。则对每一个三维点X_j来说这组观测矩阵元素可以写成式(5)。

$$\begin{pmatrix} u_{ij} \\ v_{ij} \end{pmatrix} = x_{ij} - x_{1j} \tag{5}$$

$$= \frac{f}{d_0} \begin{pmatrix} (R_{11}^i - 1)X_j + R_{12}^i Y + R_{13}^i Z_j + t_x^i \\ R_{21} X_j + (R_{22}^i - 1)Y_j + R_{23}^i Z_j + t_x^i \end{pmatrix} \tag{6}$$

则显然的在仿射投影关系下，观测矩阵也是由包含姿态的矩阵乘以包含了三维点信息的矩阵得到。

对投影相机来说，上述的简单关系更复杂一些，因为齐次化过程依赖于每个像素的深度信息[7]。考虑到齐次坐标到非齐次坐标的变换，为了能够通过矩阵分解的方式进行求解，我们把齐次化中的尺度因子作为一个需要先求解的参数进行处理。我们把投影关系下的尺度因子记做$\lambda \in \mathbb{R}$。我们记$P_i \in \mathbb{R}^{3\times 4}{}_{i=1}^{f}$ 是一系列相机的投影矩阵，包含了旋转和投影及内参的关系，把齐次坐标表示的三维点$X_j \in \mathbb{R}^{4}{}_{j=1}^{p}$映射到第i个相机里,即满足投影关系$\lambda_{ij}x_{ij} = P_iX_j$。在整个视频序列以及整个三维点集上完整的投影投影过程如式(7)所示，

$$W = \begin{pmatrix} \lambda_{11}x_{11} & \cdots & \lambda_{1p}x_{1p} \\ \vdots & \ddots & \vdots \\ \lambda_{f1}x_{f1} & \cdots & \lambda_{fp}x_{fp} \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_f \end{pmatrix} \begin{pmatrix} X_1, & \cdots, & X_p \end{pmatrix}. \tag{7}$$

Sturm 和 Trigss[7] 根据基本矩阵和极点的约束对式(7)中的λ进行了求解。由于单目视频本来就无法恢复尺度，我们可以任意选择一个深度尺度，比如设$\lambda_{1p} = 1$作为初始值。根据不同帧之间的极线约束，这些射影相机观测矩阵中的深度尺度可以按照式(8)进行更新求解：

$$\lambda_{mp} = \frac{(e_{mn} \times x_{mp}) \cdot (F_{mn}x_{np})}{\|e_{mn} \times x_{mp}\|}\lambda_{np}. \tag{8}$$

其中$m, n \in 1, 2, \cdots, f$，F_{mn}和e_{mn}分别为m对n帧之间定义的基本矩阵和极点。在求解得所有的深度尺度λ_{mp}之后我们就可以得到射影相机模型下的观测矩阵。

针对射影相机的另一个处理方式是Liu等人[6]所采用的小运动近似的方式。这种方式中假设在视频序列中相机的运动相对与帧率来说非常缓慢，每帧之间的旋转运动较小，可以使用李代数到旋转矩阵的一阶泰勒展开做为近似，在这个近似下线性关系更加明确，能够得到简单的观测矩阵。

1.4 针对多刚体系统的观测矩阵分解

我们先从静态世界或者整体就是一个刚体的系统进行考虑，通过1.2节的方式，我们可以在一个视频序列中得到它的观测矩阵$W \in \mathbb{R}^{\nvDash \mho \times |}$，这里$f$是帧数$p$是三

维点数。根据1.3节中的推导，我们可以看出这个观测矩阵可以分解成运动矩阵$M \in \mathbb{R}^{2f\times 4}$和形状矩阵$S \in \mathbb{R}^{4\times p}$即如式(9)所示：

$$W = MS. \tag{9}$$

在求解过程中，我们在得到观测矩阵W之后，基于秩约束(Rank constraint)[8]，矩阵W可以用奇异值分解（SVD）进行分解，得到

$$W = U\Sigma V, \tag{10}$$

的形式。其中$\Sigma \in \mathbb{R}^{4\times 4}$是一个对角阵，包含了最大的四个特征值，$U \in \mathbb{R}^{2f\times 4}$和$V \in \mathbb{R}^{p\times 4}$是对应到最大的四个特征值的特征向量。之后我们可以用$\hat{M} = U\Sigma^{1/2}$和$\hat{S} = \Sigma^{1/2}V^T$来表示运动矩阵和形状矩阵。但是式(10)中的分解并不是唯一的，真实的运动矩阵M和形状矩阵S还需要再找到一个映射矩阵A使得整个分解过程如下式所示：

$$W = MS = (\hat{M}A)(A^{-1}\hat{S}). \tag{11}$$

其中矩阵A可以通过旋转矩阵和平移所带有的先验约束进行求解，并可以转化成一个最小二乘形式线性求解过程[8, 1]。

上述就是通过分解形式联合求解静态场景问题的基本框架，这个框架可以比较容易的推广到场景中存在独立运动的多个刚体的情形[8]。我们假设场景中包含了n个独立运动的刚体，则我们可以通过列交换的形式把观测矩阵W中的特征序列根据刚体分组成$[W_1, \cdots, W_n]$，这个过程可以用一个排列矩阵Γ来表示，如式(12)所示，其中$\Gamma \in \mathbb{R}^{p\times p}$是一个未知的排列矩阵。

$$\bar{W} = W\Gamma = \Big(W_1, \cdots, W_n\Big) \tag{12}$$

在没有噪声的情况下，每一个独立的W_i即和前述静态场景中的观测矩阵等价应当在一个秩不超过4的子空间中。这样每个W_i可以进行单独的分解，得到该刚

体的运动矩阵M_i以及形状矩阵S_i，如下式所示：

$$\bar{W} = \bar{M}\bar{S} = \Big(M_1, \cdots, M_n\Big) \begin{pmatrix} S_1 & & \\ & \ddots & \\ & & S_n \end{pmatrix}. \tag{13}$$

这样对多刚体运动的问题来说，最关键的就是求解排列矩阵Γ让分解得到的矩阵$\bar{S}$具有块对角的性质。

多刚体运动恢复结构（Multibody Structure from Motion）对标准SfM下刚体相机的运动进行了拓展，变成了n个刚体的刚性运动模型。为了解决多刚体运动恢复结构问题，在仿射相机模型的假设下Costeira和Kanade[8] 引入了一个形状交互矩阵的概念。这个理论里对物体形状构造了一个数学上的可证明的、对刚体运动具有不变性、不依赖于坐标系的描述。这里的结构交互矩阵被证明可以保持在原始子空间里的结构。我们设$\bar{W} = U\Sigma V^T$是一个秩r的观测矩阵SVD分解结果，其中$U \in \mathbb{R}^{2f\times r}$, $\Sigma \in \mathbb{R}^{r\times r}, V \in \mathbb{R}^{p\times r}$。则形状交互矩阵$Q$就可以定义为：

$$Q = VV^T \in \mathbb{R}^{p\times p}. \tag{14}$$

式(14)具有一个特殊的形状，当两个特征序列分别属于是两个刚体的时候，Q的元素会是0。 这个特性可以在数学上得到证明[9]。 在这个理论基础上，矩阵分解的求解过程可以基于对Q的排序和元素大小的限制来或得不同刚体的分割以及三维信息的重建。 在[8]中，这种运动刚体的分割和聚类是通过最大化式(13)中S的对角元素，并且利用Q来约束对角线上的每个块应当属于不同的刚体的约束这样的方式进行求解。Ichimura[10] 使用在[11]中最大化不同子空间的差异性的判别准则将Q中的不同刚体到不同的刚体。

1.5 针对非刚性运动的观测矩阵分解

假如场景中的物体是非刚性运动的时候，情况会非常复杂。Bregler[12]等针对非刚性运动恢复结构上针对垂直投影相机提出了第一个观测矩阵分解的方式。

他们的核心想法是将一个非刚体的物体表示成一组基的组合，认为这个刚体在这些帧中的运动过程可以在这个状态空间中进行近似表示。比如我们选择k个关键帧中的结构作为一组基$B_{i i=1}^{k}$，其中每个B_i代表了一个$3 \times p$的矩阵，表示了p个特征点。这组基的线性组合$B = \sum_{i=1}^{k} l_i B_i$可以确定一个刚体的描述，其中$l_i \in \mathbb{R}$是一组系数。通过[1]中的方法进行中心化并消去平移向量之后，可以将观测矩阵表示成式(15)所示：

$$\tilde{W} = NB = \begin{pmatrix} l_{11}R'_1 & \cdots & l_{1k}R'_1 \\ \vdots & \ddots & \vdots \\ l_{f1}R'_f & \cdots & l_{fk}R'_f \end{pmatrix} \begin{pmatrix} B_1 \\ \vdots \\ B_k \end{pmatrix}. \tag{15}$$

其中R'是旋转矩阵的前两行，由于垂直投影的假设这里R就是旋转矩阵不包含内参的过程，所以旋转矩阵的第三行可以通过前两行的叉乘得到。针对观测矩阵$\tilde{W}$进行SVD分解，根据秩3的约束，我们选择3个最大的奇异值及对应的特征向量。则旋转矩阵元素R'_i和形状基的系数l_ij可以从N中，通过重新排列N的顺序并且对它进行SVD分解恢复出来。与刚性问题中面临同样的问题即SVD可以得到的结果是不唯一的，最后可以通过正交约束求解得到一个映射矩阵G得到R'_f和B_k的唯一解。另一中约束是[13]中引入了一个新的基约束，使得非刚性的分解问题能够使用闭解的形式进行求解。除了直接使用度量约束（metric constraints），Paladini等[14]将运动矩阵投影到一个矩阵流形的约束上，让整个分解过程可以通过迭代的方式进行处理。在这些工作的基础上，Dai等[15]尝试去掉针对非刚性重建的额外假设，比如之前使用的非刚性基、针对非刚性场景的先验等，提出了一个没有额外先验的仅使用低秩约束的方法。Kumar等[16]提出了融合多刚体和非刚体的方法，将问题建模成多个非刚体变换的系统。他们讲整个特征轨迹建模乘联合的多个线性或者仿射的空间。可以允许同时优化非刚性的重建和刚性的重建。

2　动态环境下的SLAM系统

3　长时变化环境下的地图更新

相比于基于稀疏特征的视觉SLAM算法，稠密视觉SLAM技术（dense visual SLAM）通过更注重于维护高质量、可复用的三维地图来帮助传感器定位。由于便携的消费级深度传感器的出现，室内场景的稠密视觉SLAM在近些年取得了不错的进展。KinectFusion [18]首次利用RGBD数据实现了实时的稠密定位和数据融合，并在场景尺度 [19]、回环调整 [20]以及计算效率 [21]上有着一系列的拓展。这类方法建立在场景完全静态的严格假设下，当运动物体区域的点云数据被融合到三维地图中，将会带来系统不可逆的崩塌。现有的针对动态场景和环境变化下的定位方法可以分为三类：一类方法只将观测数据中的静态区域融合到三维地图中，确保基于地图的定位方法仍然建立在静态世界的假设下；一类方法分别构建静态地图和动态地图，利用动态地图的历史时序信息作为先验，以提升系统的精度和鲁棒性；还有一类方法维护整个地图在时序上的变化情况，通过引入时间维度的信息将环境描述成一个随时间而转移的状态量，通过反映出的环境变化情况提供更好地预测信息。

3.1　动态环境下静态部分地图的构建

如前文所述，对于存在运动物体的场景，动态的观测数据违反了基本的几何约束，需要被视为离群点从地图中剔除，这一思想与基于运动分割的SLAM技术十分相似。对于稠密视觉SLAM任务来说，维护静态的三维地图能充分地进行数据融合，也为观测提供了更加完整的运动状态先验。相应地，应对运动物体的挑战主要包括如何避免将运动部分的数据融入地图，如果运动部分数据没有被很好地消除，用于定位的地图信息就会使问题变得复杂起来。

事实上，通过维护静态地图和采用鲁棒的定位策略在很早的时候就被广泛研究。Fox等人 [17]发现，Markov localization通过维护整个状态空间的概率密

Figure 2: 博恩德意志博物馆中的交互式解说机器人Rhino可以在人群中进行准确的自定位 [17]。

度，可以在环境偶尔变化的情况下能够保持稳定，比如门的开关或人的走动。然而，当大量物体没有包含在静态地图中，比如摄像头被室内的人群包围时(如图 ??所示)，相机定位将会失败，其主要原因在于马尔科夫假设在高动态环境下并不成立。Fox等人利用entropy filter和distance filter两种滤波方法选出输入数据中没在地图中的部分，将状态空间离散化，从而高效准确地更新置信状态，保证传感器在复杂动态场景下的定位鲁棒性。

ElasticFusion [22]可以应对画面中存在少量运动物体的场景。算法并未显式地检测运动物体，而是将动态环境下的稠密重建作为一个鲁棒估计问题，通过统计的方式自主地将动态区域作为外点剔除。在这个工作的基础上，[23] 从重建的角度出发，认为每个面元只有在多个连续帧被反复观测到才可以融合到三维模型中。当输入的点云数据与匹配上的地图点位置距离过远时，这部分点云会被作为种子点，通过区域生长将当前帧分割成静态和动态区域。相应地，地图上与动态区域有着匹配关系的部分将从地图上剔除掉。通过这种不断更新地图的方式，当之前静态的物体发生运动时，系统可以有效地检测出运动状态的变化，以消除这部分数据对系统鲁棒性的影响。

BaMVO [24] 利用背景提取领域(background subtraction)广泛使用的非参数化背景模型进行稠密视觉里程计估计。通过存储连续的4帧深度图并对齐到同一个视角，背景区域可以根据多帧对齐后的深度值差异来进行判别。这样的多帧

判别方法建立了时域上的连续性，但是由于采用帧到帧（frame-to-frame）的定位策略，BaMVO不可避免地引入了累计误差。

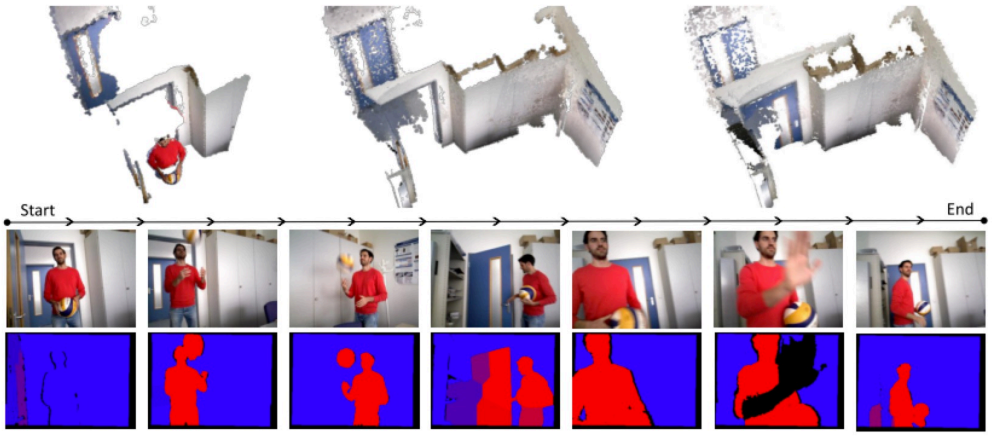


Figure 3: 三维静态地图提供了更加完整的先验，有助于提升运动分割和相机位姿估计这两个子问题的联合求解 [25]。

BaMVO说明时序多帧的反馈对动态环境下有效的运动物体检测与分割至关重要，而StaticFusion [25]认为有效的时序信息传播可以通过维护一个只包含场景中静态部分的三维地图来实现。三维数据可以有效地进行长时的三维时序信息融合，而数据融合有效地压缩了冗余信息，降低了整个系统的计算代价和内存开销。如图 ??所示，通过同时检测运动物体并重建静态环境，staticFusion实现了动态环境下的鲁棒稠密的RGBD SLAM。点云数据被聚类到一个个聚类簇中，每个聚类簇再进行运动状态估计和刚体运动估计的联合求解，以获得每个聚类簇属于静态或动态的概率。被判定为静态的聚类簇内数据会被融合到静态地图中，而被判定动态的聚类簇会进行场景流估计，以实现运动物体时序上的信息传递。由于采用了帧到模型（frame-to-model）的定位策略，相机位姿估计可以有效地消除由于累计误差带来的漂移。

DynaSLAM [26]提出了一种在线的算法，可以同时在单目、双目和RGBD相机设定下应对环境中的运动物体。整个系统建立在ORB-SLAM [27]的前端基础上，而核心出发点是通过建立可复用的三维地图进行更加精确的相机位姿估计。对

于单目相机和双目相机，DynaSLAM采用卷积神经网络（CNN）进行像素级的物体分割，作为运动状态估计的先验。在RGBD相机的设定下，由于有着更加可靠的深度信息，DynaSLAM结合了基于稠密地图的几何约束和深度学习的算法进行运动物体检测。如图 ??，通过语义信息与几何约束相结合的方式，DynaSLAM可以应对一些复杂的情形：一类是可能运动的物体在数据采集过程中处于静止状态的情形，比如停着的汽车或者坐着不动的人；另一类是没有运动先验的物体被懂地发生运动的情形，比如人推着椅子行进。这种深度学习与几何相结合的方法可以更好地应对长时复杂多变的环境，在运动状态易变的情况下尽可能地剔除可能发生运动状态变化的数据，建立更稳定可靠的静态地图来帮助定位。

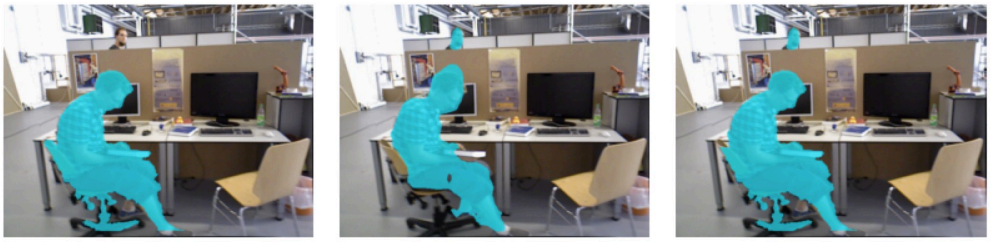


Figure 4: 将基于稠密地图的几何约束（左）和基于深度学习的语义信息（中）相结合，可以更好地在复杂的动态环境下进行运动分割（右）[26]。

当然，在动态环境中构建静态地图依赖静态世界（static world）这一基本假设。在仓库、停车场和住宅这种环境的组成容易发生变化的场景下，环境变化将持续很长的时间，而这种变化可能有利于相机的定位。在极端情形下，可见范围内的静态地图占比很少或者信息量很小的时候，对动态物体运动的推断就对相机位姿估计起到了至关重要的作用。

3.2 静态背景和动态物体的同时建图

尽管动态物体对于相机位姿的求解会造成干扰，对动态物体的运动估计对于整个系统而言仍然至关重要。一些方法将静态背景与动态物体拆分开来，分别进行三维地图的构建。相比于只维护静态地图的方法，动态物体的运动和几何结

构的推断可以带来更好的静态地图和更加可靠的运动分割结果。当然，该类方法的复杂性和计算代价也明显变高，因为算法不仅要通过识别静态背景以获得良好的位姿信息，还需对于每一个运动物体都维护独立的坐标系和地图以进行相应的位姿估计和数据融合。

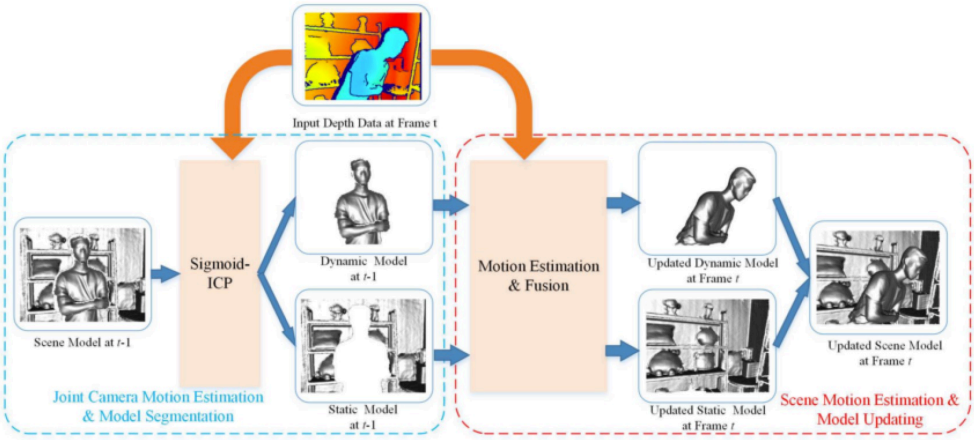
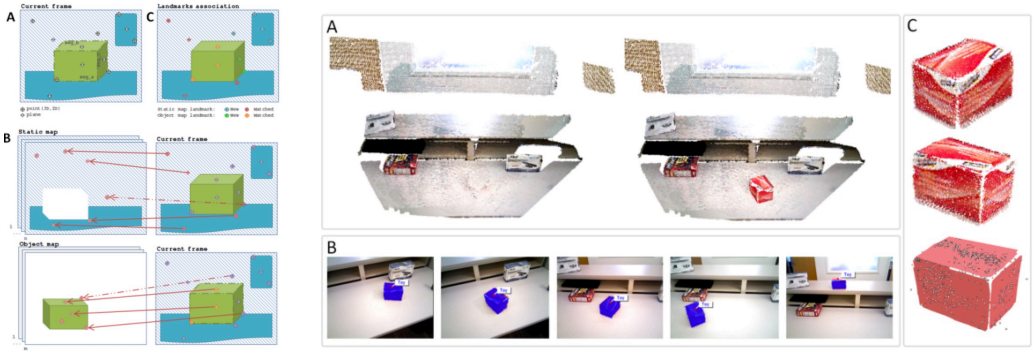


Figure 5: MixedFusion算法 [28]的流程图。蓝色虚线框内是运动分割和相机位姿估计的联合优化，而红色虚线框内是动态物体运动估计和三维数据融合。

MixedFusion [28]分别维护了每一时刻的运动物体模型和静态背景地图作为整个场景模型，并对输入的每一帧深度信息进行初步配准，区分出静态部分和动态物体。如图 5所示，场景的静态部分用于相机位姿的估计，而对于动态的部分作者则参考了Newcombe提出的DynamicFusion [29]，使用了图节点(graph Node based Motion Representation)将非刚体运动转化为以节点为控制点的多段刚体运动估计。通过运动物体的非刚体运动估计建立当前帧与模板模型（canonical model）的映射关系，来进行动态物体的数据融合。由于mixedFusion只用到了输入的深度信息而丢弃了RGB图像信息，在数据配准时容易受到深度弱纹理的影响，并且难以处理动态物体发生拓扑变化的情况。

Caccamo等人 [30]使用了自底向上（bottom-up）的特征分类的方式进行物体的识别与分割，如图 6a所示。算法首先利用第一帧数据初始化静态背景地图，然后将每一帧与静态地图进行配准。运动检测模块将特征分类并将输入数

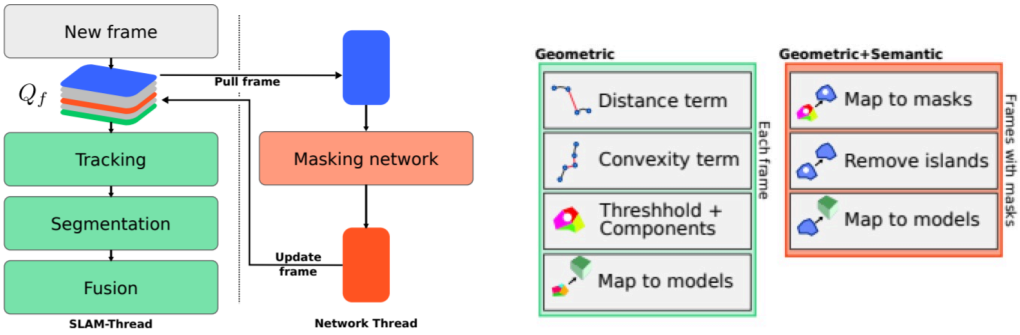
据分到维护的静态地图或物体模型。整个系统建立在基于关键帧的SLAM框架上维护了一个静态的地图，并对输入的每一帧进行特征计算与配准。根据配准之后的误差，将误差较高的部分聚合分离出来，从而判断出与相机运动不一致的动态物体，并维护该动态物体的地图，完成融合。



(a) (b)

Figure 6: （a）观测数据中的平面被组合起来，形成不同的分割区域。基于静态背景的配准可以去除由于运动造成的错误匹配。（b）真实场景下重建得到的完整场景模型。

类似的，针对多个物体的同时跟踪与场景模型重建，Rünz和Agapito [31]提出了Co-Fusion，可以处理多个不同物体的运动。该方法通过几何约束和语义信息将物体从场景中分割出来，然后对这些物体分别进行跟踪和重建。算法分割出物体后，可对每一部分的三维数据分别进行基于面元的数据融合，以处理不同物体的刚体运动，获得它们的三维模型。这种基于物体分割的动态物体重建会更适用于机器人相关的应用。算法可以对运动的物体获得较为准确的三维信息，从而使得机器人可以与环境进行更为丰富的交互。Rünz等人之后基于深度学习的方法提出了MaskFusion [32]，算法将Mask-RCNN [33]的分割结果与形状信息相结合，替代了原有的分割模块，从而在物体的分割边缘上能得到更好的表现，如图 7a所示。该类方法将语义信息与几何边缘信息相结合，从而获得更加完善的室内场景的物体分割结果。但从另一个角度来说，物体的语义信息依



(a) MaskFusion算法结构框图　　(b) 物体分割的各个约束

Figure 7: MaskFusion [33]利用二维图像的语义推断维护了场景中每个物体以及整个静态背景独立的三维模型。

赖于模型的训练集。实验过程中的运动物体需要在训练集中出现过才能得到合理的分割结果，这也是使用语义作为分割标准的一个无法避免的弊端。

相较于使用语义信息进行自顶向下的分割，Xu等人 [34]使用实例分割（instance segmentation），并通过几何和运动信息进行分割结果的优化，获得更好的分割边缘。如图8所示，三维地图中不仅仅只维护了几何和颜色信息，也保留了语义类别和运动状态的先验，以便为系统提供更加鲁棒的预测。对于分割后的物体，算法分别对这些物体进行物体姿态的估计、建图以及数据融合。由于维护了基于体素结构的物体级的三维地图，算法对环境变化和未占用空间有着更强的感知能力，在室内场景的移动机器人领域有着更广泛的应用前景。

总体而言，动态物体与静态场景的同时重建问题是一个较为困难的问题，即便输入为信息最为丰富的RGBD数据，目前也很难给出一个普适性的解决方案，均需要根据情况增加约束以使得问题可解。研究大多着眼于如何区分静态与动态部分，并使用适当的模型来描述动态物体的运动。尽管目前对于单一物体的简单运动可以恢复出较好的模型，但对于多物体复杂运动，考虑到相应的运算开销，常常难以获得较为鲁棒、准确的结果。另一方面，虽然运动部分的输入数据也被保留在地图中，这类方法仍然遵循着静态世界的基本假设，在动态变

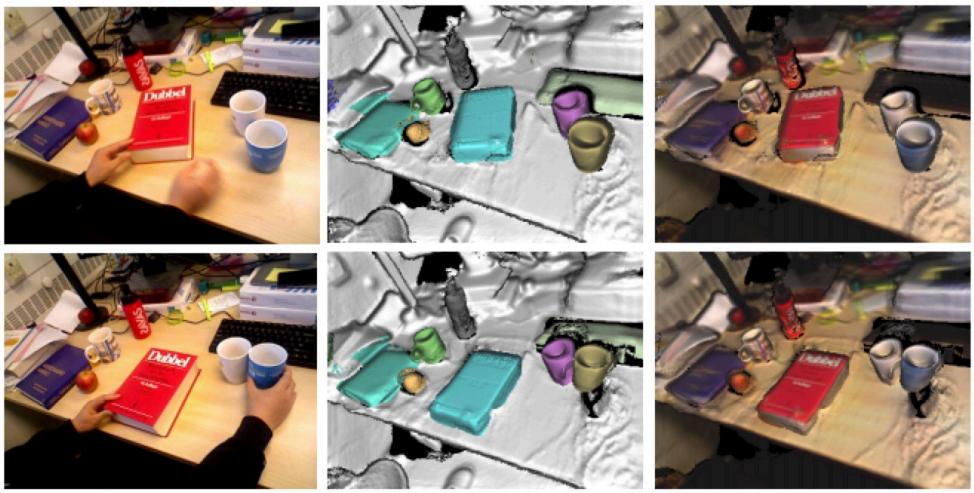


Figure 8: MidFusion算法 [34]构建了物体级（object-level）的稠密体素地图，可以应对移动的物体，并忽略场景中人的运动。

化较弱的环境下，同时维护静态地图和动态物体模型仍然会和只保留静态地图的方法遇到相似的挑战。

3.3 四维地图构建与长时定位

对于动态场景建模，先前的一部分方法完全将动态区域作为离群点予以剔除，另一部分方法则同时维护静态和动态地图，以提供一个更好的环境静态地图和更可靠的动态区域检测。但无论哪种途径，都依赖于静态世界的假设，这使得这些方法在部署到不断变化的环境或是动态性较低的环境中时效果不佳。

为了克服静态世界假设的局限性，一些研究人员致力于在一个统一的表示当中建模环境的动态性，并最终达到进行lifelong 建图的目的。Chen等人[35]以及后来的Brechtel等人[36] 提出并拓展了传统的占据网格的框架，使之包含了对动态物体的建模，并用贝叶斯滤波的方式对其进行更新。在这个视角下，针对动态性他们建议了以物体为中心的表示，他们认为占据网格中格子的占据概率由环境中的物体决定，当物体发生运动，其对应的占据网格也会发生相应的

运动。因此，在该框架中，他们需要自始自终追踪每一个网格的运动。与该思想相反的，采用以地图为中心的方法也可以对环境中的动态进行建模。

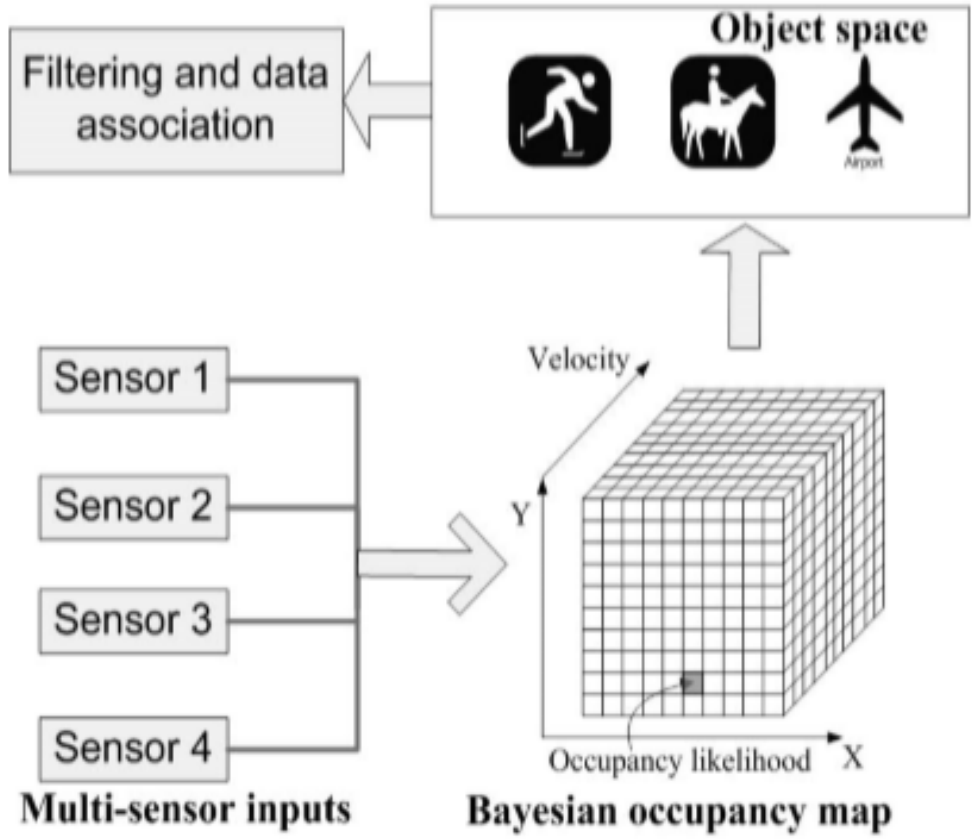


Figure 9: 基于贝叶斯的占据网格框架。

Schindler和Dellaert等人[37]利用自下而上的启发式方法，将从SfM管道中的点观测分组为建筑假设和概率时间模型来推断建筑物存在的时间间隔，建立了一个“4D城市”模型。

Yang和Wang[38]建议用一个“可能性”网格来同时表示静态区域和动态区域。一对对偶传感器模型被用来在移动机器人定位中判别静态及动态物体，然而，他们的工作假定机器人的位置是已知的，具有一定的精度来进行计算以及更新地图，所以该方法并不适合于全局定位问题。

之后，Saarinen[39]等人提出用一系列独立的马尔科夫链去建模整个环境，

Figure 10: “4D城市”示意图（1956年-1971年）。

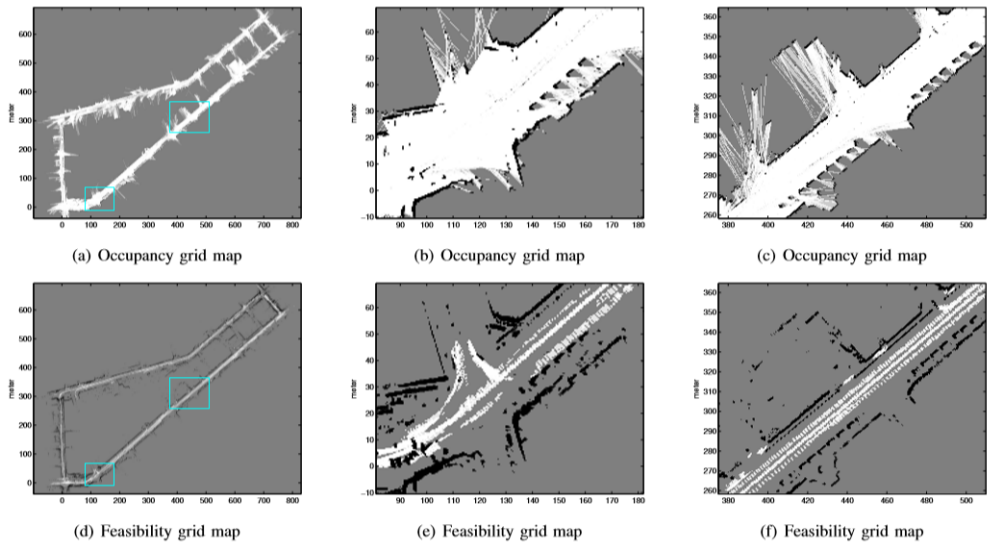


Figure 11: 含“可能性”的网格图。

将状态之间的转换参数建模为两个泊松过程并在线学习这些参数，采用基于近因加权的方法处理非平稳单元的动力学问题。而同样的，该方法也无法普适地应对真实环境下的不同的动态场景及物体。

Murphy[40]等人建议应用Rao-Blackwellized粒子滤波器来解决SLAM问题并理论上展示了其在动态场景下的可行性。但他们的方法假设了状态转换的概率与环境的当前状态独立并且给定了一个先验，且只能在一个小尺度的环境下工作。之后，Avots等人[41]，Petrovskaya[42]等人分别提出对它的改进，前者用Rao-Blackwellized粒子滤波器来估计机器人的姿态和环境中门的状态，他们使用一个参考占用网格来表示环境，而非他们的状态（其中门的位置是已知的）；后者与前者相似，但将门的开关状态这一二元模型改为一个参数化模型（门的打开角度）。而Stachniss和Burgard[43]也使用Rao-Blackwellized粒子滤波器对聚类后的局部网格图确定的一组可能的环境配置来对机器人进行定位，并从该集合中估计环境的配置。Meyer和Delius[44]跟踪那些由环境中使用临时局部地图的离群对象引起的观测结果，然后用上述粒子滤波器来估计机器人的姿势，该滤波器不仅依赖于这些临时地图，也依赖于环境的参考地图，然而，这项工作仍然依赖于全局定位的静态映射，只有在位置跟踪失败时才会创建临时映射。

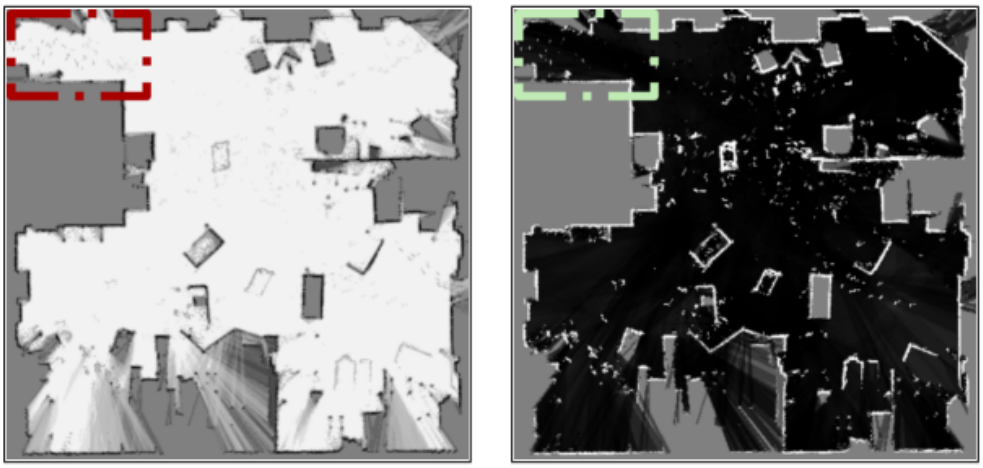
另外，对于lifelong的动态环境建图，Konolige[45]提出了一个有趣的方法，该方法主要侧重于可视化地图，并提供了一个框架，在该框架中，可以随着时间的推移更新本地地图（视图），并在环境配置更改时添加/删除新的本地地图。Kretzschmar[46] 等人也给出了类似的想法，他们利用一种有效的信息论图形修剪策略进行图形压缩。该方法可用于偏倚最近的观察结果，以获得与前者工作的类似的表现。然而，这两种方法主要集中在长期操作中出现的可伸缩性问题上，而不是环境随时间变化的动态方面。从这个想法出发，Walcott-Bryant[47]等人提出了一个名为Dynamic Pose Graph （DPG）的局部表示来建模长时下低动态环境的SLAM问题。

Churchill和Newman[48]提出了关于lifelong建图的另一个视角。他们认为

导航不需要一个全局参考框架，并介绍了“经验”的概念，即具有相对测量信息的机器人路径。“经验”可以通过基于外观的数据关联方法连接在一起，随着时间的推移而变化的地方由一组不同的“经验”表示。 Tipaldi等人[49]改进并综合了上述基于粒子滤波的方法，提出了一种新的适应环境变化的lifelong 定位方法，它明确地考虑了环境的动态变化，且能够区分表现出高动态行为的物体，例如汽车和人，可以移动并改变配置的物体，例如箱子、架子或门，以及静止不移动的物体，例如墙壁。该方法在二维网格上用一个隐马尔科夫模型描述空间的占据和它的动态性，并通过EM算法学习其参数，联合估计机器人姿态以及全局定位中的环境状态，然后应用一个Rao-Blackwellized粒子滤波器（其中机器人姿态为被采样部分滤波器，网格占据状态为分解的解析部分），同时通过考虑相关马尔可夫链的混合时间来建立一种基于局部地图表示的地图管理方法以能够最小化内存需求，并以合理的概率方式来忘记变化。

之后，Krajník等人[50]提出在光谱域中表示环境动力学，并将其用去图像特征以改进定位，之后也陆续有研究者将该方法应用于占用网格以减少内存需求、应用于拓扑图以改进路径规划。

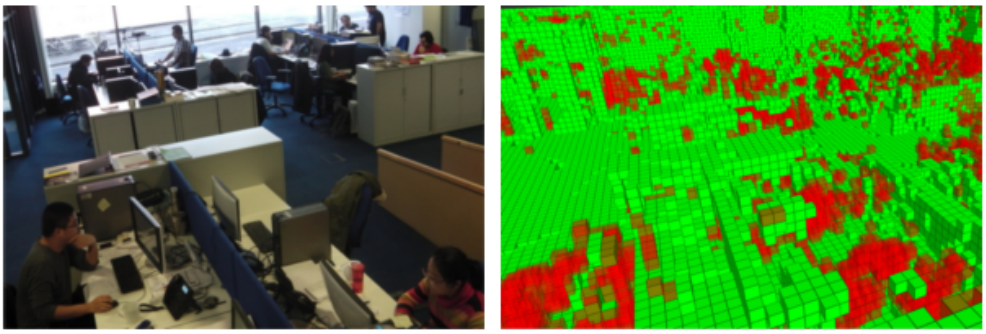
虽然上述方法适用于移动机器人中使用的大多数环境模型，但由于其依赖于传统的快速傅立叶变换（FFT）方法，因此存在一个主要缺陷，即需要对环境进行定期和定期的观测。这意味着机器人的活动必须分为一个学习阶段，当它经常访问各个位置建立其动态环境模型时，以及当它使用其模型执行有用任务时的部署阶段。这一划分意味着，虽然机器人可以创建更适合长期操作的动态模型，但它不能维护这些模型。因此，机器人不适应那些不存在学习阶段的动力学问题，这会导致其效率随着时间的推移而降低。Krajník等人[51]又提出了一种lifelong移动机器人时空动态环境探测的新思路，该方法假设世界处于不断变化的状态，这将为探索空间增加一个额外的时间维度，使探索任务成为一个永无止境的数据收集过程。为了创建和维护一个动态环境的时空模型，机器人不仅要确定在哪里，还要确定何时进行观察。我们将信息论探索应用于世界表征，将环境状态的不确定性建模为时间的概率函数，从而解决这一问题。



(a) $p_c^{o|f}$ (b) $p_c^{f|o}$

(c) $p_c^{o|f}$ (d) $p_c^{f|o}$

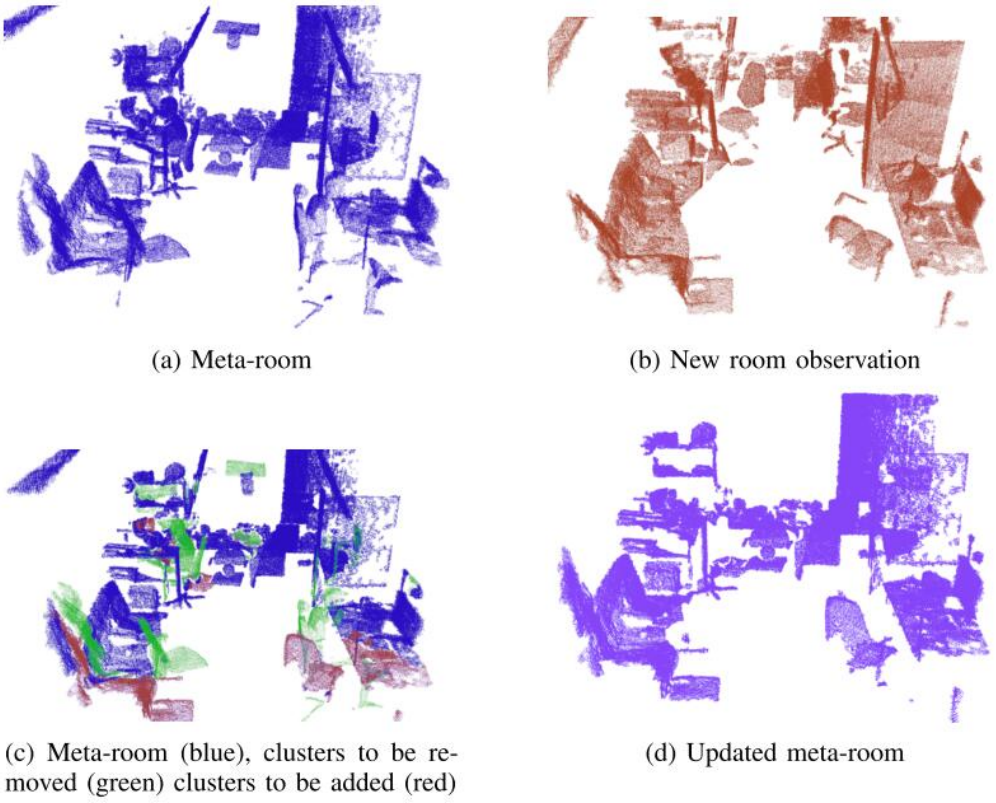
Figure 12: 状态转移概率示意（颜色越深，概率越大）。



(a) Brayford office view (b) Brayford spatio-temporal model

Figure 13: 随时间变化的占据网格（红色部分）。

另外，Ambrus等人[52]提出了一种新的方法来重新创建杂乱的办公环境的静态结构，他们将其定义为“meta-room”，它基于一个配备了rgb-d 深度摄像头的自主机器人在长时间内收集到的多个观测结果进行实验。该方法通过识别从一个观测点到下一个观测点的变化，移除动态元素，同时添加先前被遮挡的对象，以尽可能准确地重建底层静态结构，直接与点簇一起工作。构建meta-room的过程是迭代的，它被设计为在可用时合并新数据，并对环境变化具有鲁棒性。meta-room的最新估计用于区分和提取动态物体群与观测结果。该方法之后也被应用在一些导航机器人平台来得到更好，更细节的物体模型。



(a) Meta-room

(b) New room observation

(c) Meta-room (blue), clusters to be removed (green) clusters to be added (red)

(d) Updated meta-room

Figure 14: meta-room更新过程示意。

上面提到的Krajnik和Ambrus的工作重点都是使用变化检测算法的结果来分析变化的时空行为，而有的研究人员只关注观测结果之间的变化。Fehr等人[53]提出了一种新的基于扩展截断有符号距离函数（TSDF）的动态场景下的三

维重建算法，该算法能够在场景中同时获得动态对象的三维重建的同时，对静态地图进行连续的细化。这是一个具有挑战性的问题，因为地图更新是递增的，并且常常是不完整的。以前的工作通常在点云、曲面或地图上执行变化检测，这些点云、曲面或地图无法区分未探测空间和空白空间。相比之下，该方法基于TSDF的表示自然包含了这些信息，从而使其能够更有力地解决场景差异问题。

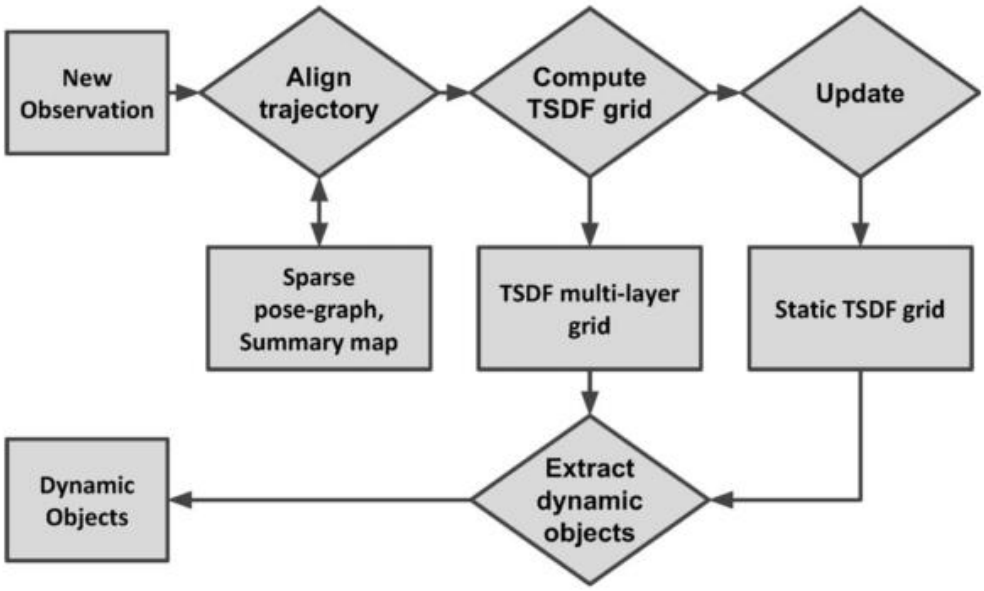


Figure 15: 基于TSDF的变化检测框架。

总而言之，关于如何将静态和动态场景置于一个统一优美的空间表示形式下，早期研究人员在基于滤波的框架下作了很多的探索，而在面对实际问题时，大部分在真实场景下拥有鲁棒效果的方法却仍然需要沿着前面几个章节所述的技术路线进行。

References

[1] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. Intl. J. of Computer

Vision, 9(2):137－154, 1992.

[2] Nolang Fanani, Matthias Ochs, Henry Bradler, and Rudolf Mester. Keypoint trajectory estimation using propagation based tracking. In IEEE Intelligent Vehicles Symposium (IV), 2016.

[3] João Costeira and Takeo Kanade. A multi-body factorization method for motion analysis. In Intl. Conf. on Computer Vision (ICCV), 1995.

[4] Michal Irani. Multi-frame correspondence estimation using subspace constraints. Intl. J. of Computer Vision, 48(3):173－194, 2002.

[5] Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. Subspace video stabilization. ACM Trans. Graphics, 30(1): 1－10, 2011.

[6] Liu Feng, Yuzhen Niu, and Hailin Jin. Joint subspace stabilization for stereoscopic video. In Intl. Conf. on Computer Vision (ICCV), 2013.

[7] Peter Sturm and Bill Triggs. A factorization based algorithm for multi-image projective structure and motion. In European Conf. on Computer Vision (ECCV), pages 709－720, 1996.

[8] João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. Intl. J. of Computer Vision, 29(3):159－179, 1998.

[9] Ken-ichi Kanatani. Motion segmentation by subspace separation and model selection. In Proceedings Eighth IEEE International

Conference on computer Vision. ICCV 2001, volume 2, pages 586-591. IEEE, 2001.

[10] Naoyuki Ichimura. Motion segmentation based on factorization method and discriminant criterion. In Intl. Conf. on Computer Vision (ICCV), 1999.

[11] N. Ostu. A threshold selection method from gray-histogram. 9(1): 62-66, 2007.

[12] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.

[13] Jing Xiao, Jinxiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. Intl. J. of Computer Vision, 67(2):233-246, 2006.

[14] Marco Paladini, Alessio Del Bue, Marko Stosic, Marija Dodig, João M. F. Xavier, and Lourdes De Agapito. Factorization for non-rigid and articulated structure using metric projections. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.

[15] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012.

[16] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Multi-body non-rigid structure-from-motion. 2016.

[17] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Markov localization for mobile robots in dynamic environments. Journal of artificial intelligence research, 11:391–427, 1999.

[18] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR), volume 11, pages 127–136, 2011.

[19] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. ACM Trans. Graphics, 32(6):169, 2013.

[20] T Whelan, M Kaess, MF Fallon, H Johannsson, JJ Leonard, and JBM Kintinuous. Kintinuous: Spatially extended kinectfusion. In RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, 2012.

[21] Frank Steinbrücker, Jürgen Sturm, and Daniel Cremers. Volumetric 3d mapping in real-time on a cpu. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 2021–2028. IEEE, 2014.

[22] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. Robotics: Science and Systems (RSS), 2015.

[23] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In International Conference on 3D Vision (3DV), pages 1–8. IEEE, 2013.

[24] Deok-Hwa Kim and Jong-Hwan Kim. Effective background model-based rgb-d dense visual odometry in a dynamic environment. IEEE Trans. Robotics, 32(6):1565－1573, 2016.

[25] Raluca Scona, Mariano Jaimez, Yvan R Petillot, Maurice Fallon, and Daniel Cremers. Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 1－9. IEEE, 2018.

[26] Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. IEEE Robotics and Automation Letters, 3(4):4076－4083, 2018.

[27] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Trans. Robotics, 33(5):1255－1262, 2017.

[28] Hao Zhang and Feng Xu. Mixedfusion: Real-time reconstruction of an indoor scene with dynamic objects. IEEE Trans. on visualization and computer graphics, 24(12):3137－3146, 2017.

[29] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 343－352, 2015.

[30] Sergio Caccamo, Esra Ataer-Cansizoglu, and Yuichi Taguchi. Joint 3d reconstruction of a static scene and moving objects. In International Conference on 3D Vision (3DV), pages 677－685, 2017.

[31] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 4471－4478, 2017.

[32] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR), pages 10－20, 2018.

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Intl. Conf. on Computer Vision (ICCV), pages 2961－2969, 2017.

[34] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 5231－5237, 2019.

[35] C. Chen, C. Tay, C. Laugier, and K. Mekhnacha. Dynamic environment modeling with gridmap: A multiple-object tracking application. In International Conference on Control, 2006.

[36] Sebastian Brechtel, Tobias Gindele, and Rudiger Dillmann. Recursive importance sampling for efficient grid-based occupancy filtering in dynamic environments. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2010.

[37] Grant Schindler and Frank Dellaert. Probabilistic temporal inference on reconstructed 3d scenes. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.

[38] Shao Wen Yang and Chieh Chih Wang. Feasibility grids for localization and mapping in crowded urban scenes. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2011.

[39] Jari Saarinen, Henrik Andreasson, and Achim J. Lilienthal. Independent markov chain occupancy grid maps for representation of dynamic environments. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2012.

[40] K. Murphy. Bayesian map learning in dynamic environments. In Advances in Neural Information Processing Systems (NIPS), 1999.

[41] Dzintars Avots, Edward Lim, Romain Thibaux, and Sebastian Thrun. A probabilistic technique for simultaneous localization and door state estimation with mobile robots in dynamic environments. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2002.

[42] Anna Petrovskaya and Andrew Y. Ng. Probabilistic mobile manipulation in dynamic environments, with application to opening doors. In International Joint Conference on Artifical Intelligence, 2007.

[43] Cyrill Stachniss and Wolfram Burgard. Mobile robot mapping and localization in non-static environments. In National Conference on Artificial Intelligence, 2005.

[44] D. Meyer-Delius, J. Hess, G. Grisetti, and W. Burgard. Temporary maps for robust localization in semi. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2010.

[45] Kurt Konolige and James Bowman. Towards lifelong visual maps. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2009.

[46] Henrik Kretzschmar and Cyrill Stachniss. Information-theoretic compression of pose graphs for laser-based slam. Intl. J. of Robotics Research, 31(11):1219–1230, 2012.

[47] A. Walcott-Bryant, M. Kaess, H. Johannsson, and J. J. Leonard. Dynamic pose graph slam: Long-term mapping in low dynamic environments. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2012.

[48] Winston Churchill and Paul Newman. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2012.

[49] Gian Diego Tipaldi, Daniel Meyer-Delius, and Wolfram Burgard. Lifelong Localization in Changing Environments. 2013.

[50] Tomš¢? Krajnšªk, Jaime Pulido Fentanes, Grzegorz Cielniak, Christian Dondrup, and Tom Duckett. Spectral analysis for long-term robotic mapping. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2014.

[51] Tomas Krajnik, Joao M. Santos, and Tom Duckett. Life-long spatio-temporal exploration of dynamic environments. In European Conference on Mobile Robots, 2015.

[52] Rares Ambrus, Nils Bore, John Folkesson, and Patric Jensfelt. Meta-rooms: Building and maintaining long term spatial models in

a dynamic world. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2014.

[53] Ivan Dryanovski J¡ § urgen Sturm Igor Gilitschenski Roland Siegwart Marius Fehr, Fadri Furrer and Cesar Cadena. Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2017.