ClusterVO

- 论文：ClusterVO Clustering Moving Instances and Estimating Visual Odometry.pdf
- 开源：来自作者的主页https://drive.google.com/file/d/1zQr11Ne_52HTXcIHRH5rsbfuhi4B8HJZ/view?usp=sharing

介绍

提出一种双目视觉里程计ClusterVO，能同时完成对自我和周围刚性簇/物体的运动进行聚类与估计。

1) 本文提出的 Cluster VO 同时优化相机位姿和多个动态目标的姿态（将其视为点路标的簇），实现良好的跟踪和分割性能，此外没有任何几何或形状先验，展现了系统的通用性。本文策略仅基于稀疏路标和 YOLO 2D 目标检测。

2) 提出一种鲁棒的多级概率数据关联技术，以有效地跟踪 3D 空间中的低层特征和高层检测。

3) 在此基础上，结合语义bbox、空间信息和运动一致性，提出了一种高效的异构 CRF 算法，用于发现新的聚类、聚类新的标记和细化已有的聚类。

4) 最后，采用滑动窗口优化的方法对场景的静态部分和动态部分进行求解。

系统框架

ClusterVO 将经过同步和校准的**立体图像作为输入，并为每帧输出相机和物体位姿**。对于每个传入帧，使用 YOLO 检测语义边界框，并提取 ORB 特征在整个立体图像上进行匹配。 **1）**首先通过多级概率关联公式将检测到的边界框和提取的特征分别关联到先前发现的簇(已经聚类得到的簇)和地标。**2）**然后，我们对所有具有相关地图路标的特征执行异构条件随机场（CRF），以确定当前帧的聚类分割。 **3）**最后，状态估计步骤通过边缘化和平滑运动先验优化滑动窗口上的所有状态。该流程下图所示。

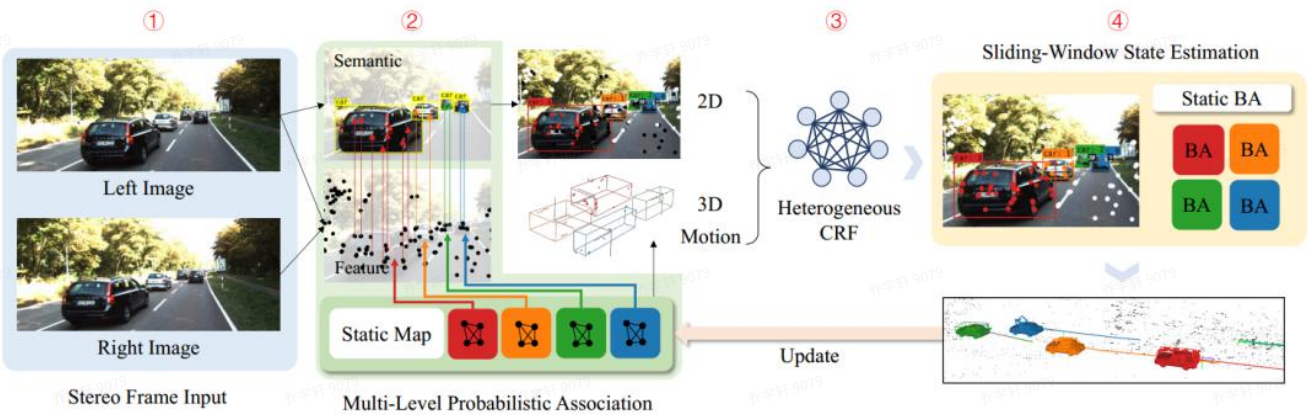


图 1：系统框架

1.1 多级概率关联

对于 **静态地图的路标**，可以通过 **最近邻搜索和描述符匹配**来稳健地关联特征 [ORB-SLAM2]。然而，在图像空间上 **快速移动的动态地标**跟踪并不是一件容易的事情。此外，如果可能的话，需要将每个检测到的边界框 B^m_t关联到一个现有的地图中的一个聚类，这在随后的异构 CRF 模块中是必需的。

为此，本文提出一个用于 **动态路标的多级概率关联方法**，将低层特征 z^k_t 分配源路标 ID：k→i，将高层边界框 B^m_t 分配聚类 m→q。 **概率方法的本质是用均值为 p^i_t 的高斯分布和协方差 sum_{t}^{i} 来模拟地标的位置，并考虑整个匹配过程中的不确定性。**

理想情况下，应该从 **最后一个状态估计步骤的系统信息矩阵**中提取 sum_{t}^{i}，但是计算量很大，因此，本文用最小的行列式近似 sum_{t}^{i} 变换为 Z Σ t'<ti （公式 1），可以被逐步更新，其中 R_t^c 是 P_t^c 的旋转部分。

$$\Sigma_t^i := \mathbf{R}_{t'}^{\mathbf{c}} \, {}_{\boldsymbol{Z}}\Sigma_{t'}^i \mathbf{R}_{t'}^{\mathbf{c}\top}, \quad t' := \underset{t'<t}{\operatorname{argmin}} \left| {}_{\boldsymbol{Z}}\Sigma_{t'}^i \right|, \qquad (1)$$

对于每个新帧，我们使用 v_t^q **对每个聚类执行运动预测**。将预测的 3D 路标的位置及其噪声协方差矩阵重新投影回当前帧中。 **将第 k 个观察值分配给路标 i 的概率得分为**公式（2）。对于每个观测值 k，我们 **选择与其对应的具有最高得分的路标 i 关联**。实际上公式（2）仅在 z_t^k 的一个小邻域上进行评估。

$$p_i(k) \propto \left[\|\boldsymbol{\zeta}_t^i - \boldsymbol{z}_t^k\|_{\Gamma_t^i}^2 < \gamma \right] \cdot s_{ik}, \qquad (2)$$

路标i与观测值z^k的描述符相似性

[] 指标函数

取为4.0

$$\boldsymbol{\zeta}_t^i = \pi(\boldsymbol{p}_t^i + \boldsymbol{v}_t^{\mathbf{q}}) \qquad \Gamma_t^i = \mathbf{J}_\pi \Sigma_t^i \mathbf{J}_\pi^\top$$

通过公式（3）计算香农交叉熵进一步测量关联的不确定性。其中 p_q(m)是 **第 m 个 boundingbox 分配给第 q 个聚类的概率**，如果 ε_t^q 小于 1，则认为这是一个 **成功的高级别关联**，在这种情况下 **在边界框内执行额外的暴力低级别特征描述符匹配，以找到更多的特征对应。**

$$\mathcal{E}_t^{\mathbf{q}} := -\sum_m p_{\mathbf{q}}(m) \log p_{\mathbf{q}}(m),$$

$$p_{\mathbf{q}}(m) \propto \sum_{\boldsymbol{\zeta}_t^k \in \boldsymbol{B}_t^m} (1/|\Gamma_t^i|), \qquad (3)$$

将第m个box分配给聚类q的概率

1.2 用于聚类簇分配的异构 CRF

在这一步确定 **当前帧观测到的每个路标 i 被分配的聚类簇 q^i**，应用了 **将语义，空间和运动信息结合在一起的条件随机场模型**（称为“异构CRF”），从而将公式（4）的能量降至最低。

$$E(\{\mathbf{q}^i\}_i) := \sum_i \psi_u(\mathbf{q}^i) + \alpha \sum_{i<j} \psi_p(\mathbf{q}^i, \mathbf{q}^j), \qquad (4)$$

一元势能　平衡因子　成对势能

其中一元能量决定观察到的地标 i 属于特定聚类簇 q^i 的概率。 单个 2D 项只考虑 2D 语义检测，可能在边界框的边缘附近包含许多离群点。通过添加 3D 项，可以修剪属于较远背景的路标点。然而，接近 3D 边界的特征，例如在移动车辆附近的地面上，仍然具有属于该簇的高概率，其置信度由运动项进一步细化。我们由一图来解释一元势能的三项带来的影响：

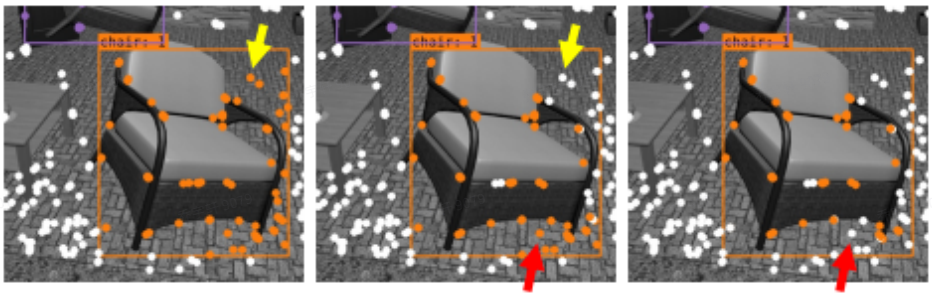


Figure 8. Unary term visualizations on one indoor sequence from SUNCG dataset. (a) ClusterVO 2D; (b) ClusterVO 2D+3D; (c) ClusterVO Full.

图(a)中，2D项将bbox内所有特征点归成一簇。图(b)中，3D项将三维坐标下远离物体(原有簇)中心的点给剔除(需要双目信息)。图(c)中，运动项引入连续帧下的运动估计，实现了物体-物体的分离(椅子和地板)。

成对势能定义为公式（9），可以 **看作是一个噪声敏感的高斯平滑核，以鼓励空间标记的连续性**。其中指数算子内的项是 3D 空间中两个路标的距离。表示两个landmark离得越近，越有可能是同一cluster。

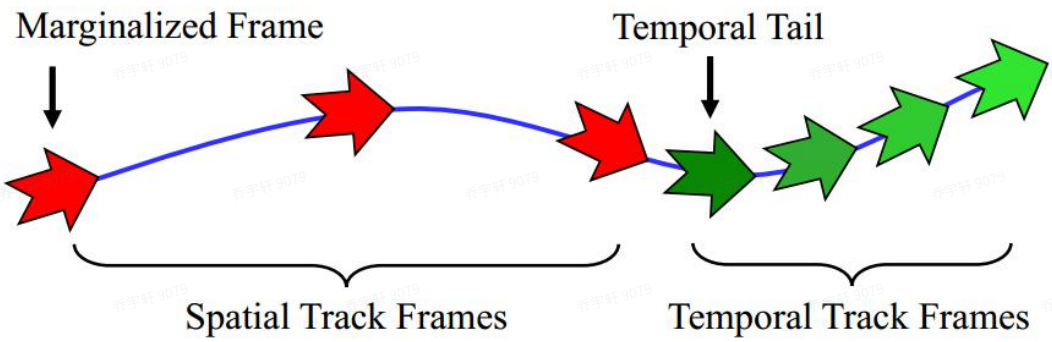
$$\psi_p(\mathbf{q}^i, \mathbf{q}^j) := [\mathbf{q}^i \neq \mathbf{q}^j] \cdot \exp(-\|\boldsymbol{p}_t^i - \boldsymbol{p}_t^j\|^2), \qquad (9)$$

本文使用一个参考文献中有效的稠密 CRF 推理方法来解决能量最小化问题。在成功的推理之后，我们得到两种有可能不同的聚类结果，对于这样的匹配问题，我们使用 Kuhn-Munkres 算法将当前的 CRF 聚类结果与之前的聚类分配相匹配。如果没有为推断的标签找到正确的群集分配，则会创建新的群集。然后，根据文献中中介绍的策略更新每个路标的权重,并在必要时更改其群集分配：当新分配的群集与路标的上一个群集相同时，将权重增加 1，否则权重将减少 1。当减少到0时，将触发集群分配的更改以接受当前分配的集群。

1.3 滑动窗口状态估计

本文根据一种新颖的 **双轨帧管理设计**（图 3）采用了滑动窗口优化方案。系统维护和优化的帧分为两个连续的轨道： **时间轨道 T_t 和空间轨道 T_s**。T_t 包含最新的输入帧，每当出现新帧时， T_t 中最旧的帧将被移出。如果该帧在空间上距离 T_s 中的第一帧足够远，或者可见的路标数量足够少，则将该帧追加到 T_s 的尾部，否则将丢弃该帧。

这种设计的 **优点**。 **1) 时间轨迹中的帧记录了所有最近的观测值**，因此允许足够的观测值来跟踪快速移动的聚类。 **2) 可以更正先前错误聚集的路标**，并可以根据新的分配进行重新优化。 **3) 在空间轨迹中检测到的特征有助于创建足够的视差**，以进行精确的路标三角化和状态估计。



静态场景和相机位姿估计，用于优化的能量函数是标准的 BA，并 **增加了一个边缘化项**。

$$\mathbf{E}(\{\boldsymbol{x}_t^{\mathbf{c}}, \boldsymbol{x}_t^{\mathbf{L}}\}_{t\in\mathcal{T}_a}) := \sum_{i\in\mathcal{I}_0, t\in\mathcal{T}_a} \underbrace{\rho(\|\boldsymbol{z}_t^i - \pi((\mathbf{P}_t^{\mathrm{c}})^{-1}\boldsymbol{p}_t^i)\|^2_{\boldsymbol{z}\Sigma})}_{\text{重投影误差项}} + \sum_{t\in\mathcal{T}_a} \underbrace{\|\delta\boldsymbol{x}_t^{\mathbf{c}} - \mathbf{H}^{-1}\boldsymbol{\beta}\|^2_{\mathbf{H}}}_{\text{边缘化项}}, \quad (10)$$

聚类q中所有路标 $\mathcal{I}_{\mathbf{q}} = \{i|\mathbf{q}^i = \mathbf{q}\}$

$\mathcal{T}_a := \mathcal{T}_s \cup \mathcal{T}_t$ 双通道的关键帧

由于静态场景涉及大量变量，并且仅将这些变量从滑动窗口中删除会导致信息丢失，从而可能产生漂移，因此需要**将某些变量边缘化**，或者将其删除，并在公式（10）中使用边缘化项表示对系统的影响。仅当从空间轨道 T_s 丢弃帧时才执行边缘化。为了限制信息矩阵中路标块的稠密填充，如果相应的路标被最新的帧所观察到，则将删除的帧中观察结果也删除。 **这种边缘化策略仅在帧中添加了稠密的 Hessian 块，而不是路标，从而使系统仍然可以实时解决(**边缘化后保持了稀疏性**)。**

关于滑动窗口边缘化可参考《视觉SLAM十四讲》P267~P269

delta_x是边缘化发生时相对于临界状态x^star的变化，由标准Schur补计算H与beta：

$$\mathbf{H} = \boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}, \boldsymbol{\beta} = \boldsymbol{b}_a - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{b}_b,$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_a \\ \boldsymbol{b}_b \end{bmatrix}.$$

动态聚类。利用加速度先验值上的白噪声对运动进行建模，在连续时间内可以写成：

$$\ddot{\boldsymbol{t}}^{\mathbf{q}}(t) \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{Q}\delta(t - t')), \quad (12)$$

聚类簇的加速度　高斯过程　功率谱矩阵

定义用于**优化**第 q 个聚类簇的**轨迹**及其对应路标**位置**的能量函数：

$$\mathbf{E}(\{\boldsymbol{x}_t^{\mathbf{q}}, \boldsymbol{x}_t^{\mathbf{L}}\}_{t\in\mathcal{T}_t}) := \sum_{t,t^+\in\mathcal{T}_t} \left\| \begin{bmatrix} \boldsymbol{t}_{t^+}^i \\ \boldsymbol{v}_{t^+}^i \end{bmatrix} - \mathbf{A} \begin{bmatrix} \boldsymbol{t}_t^i \\ \boldsymbol{v}_t^i \end{bmatrix} \right\|_{\hat{\boldsymbol{Q}}}^2 + \sum_{i\in\mathcal{I}_{\mathbf{q}}, t\in\mathcal{T}_t} \rho(\|\boldsymbol{z}_t^i - \pi(\mathbf{T}_t^{\mathbf{cq}^i}(\mathbf{P}_t^{\mathbf{c}})^{-1}\boldsymbol{p}_t^i)\|_{\boldsymbol{z}\Sigma}^2), \quad (13)$$

动态场景只优化了时间轨道　1. 运动先验误差　2. 重投影误差

$$\mathbf{A} := \begin{bmatrix} \mathbf{I} & \Delta t\mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \hat{\boldsymbol{Q}}^{-1} := \begin{bmatrix} 12/\Delta t^3 & -6/\Delta t^2 \\ -6/\Delta t^2 & 4/\Delta t \end{bmatrix} \otimes \boldsymbol{Q}^{-1}, \quad (14)$$

问题

- 在异构CRF之前，多级概率关联已经要求有簇指派，但文章对路标簇的生成并未解释。猜测路标聚类过程与作者先前文章ClusterSLAM中的聚类方法一致，主要是：给出相似度评价体系(点之间的距离度量)，在层次聚类的基础上使用共识(分块)聚类