SLAM for Highly Dynamic Environments用于高动态环境的单目语义SLAM

[TOC]

摘要

单目SLAM的最新进展使得实时系统能够在静态环境下稳定运行，但由于缺乏显式的动态离群值处理，在动态场景变化和运动时无法正常运行。我们提出了一个语义单目SLAM框架，旨在处理高度动态的环境，结合基于特征和直接的方法来实现在具有挑战性的条件下的鲁棒性。该方法利用了在显式概率模型中从场景中提取的语义信息，最大限度地提高了跟踪和建图的概率，从而依赖于那些没有相对于摄像机呈现相对运动的场景部分。我们在动态环境中显示了更稳定的姿态估计，并在虚拟KITTI和Svnthia数据集上显示了与现有技术在静态序列上的性能相当的性能。

I.介绍

在过去的几年里，在单眼同步定位和地图(SLAM)领域的激烈研究活动使其达到了看不见的准确性、鲁棒性和速度，使其在机器人和增强现实领域的各种新应用成为可能。与基于立体声或rgb - d的技术相比，单目SLAM算法[1]、[2]、[3]、[4]基于更廉价的硬件，校准更简单，深度范围没有限制，这使得它们对许多专注于户外和室内场景的移动应用程序特别有吸引力。

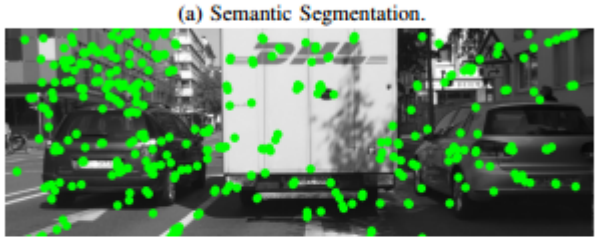
单目SLAM方法可分为两种。特征法[5]，[1]使用一个显式的关键点描述符来查找不同图像中的特征匹配，并最小化它们之间的重投影误差。与之不同的是，直接方法[6]、[2]、[4]、[3]根据像素强度从一幅图像到另一幅图像的投影来最小化光度误差。正如[3]中分析的那样，描述性方法和直接方法各有优缺点。具体来说，描述方法对几何噪声(如像素位置位移)的抵抗能力更强，而直接方法更适合处理运动模糊引起的光度量噪声。

然而，目前的单目SLAM算法依赖于周围环境是静态的假设，限制了它们对大多数现实场景的适用性。为了处理动态对象，它们要么使用m估计器

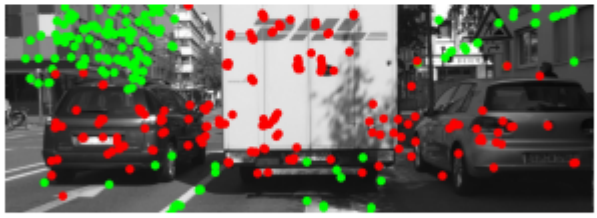
为了处理动态对象，它们要么在优化过程中使用M-estimators(Tukey[5]、Huber[2]、[1]、[3])，要么使用基于ransac的方法来检测和过滤运动[7]。为了正确工作，这两种方法都要求相对于摄像机运动的大多数点是静态的。相反，**当动态对象覆盖相机视场的主要部分时，特别是当大多数视觉特征位于这些区域时，当前的单目SLAM方法将会失败**。对于大多数与户外驾驶相关的场景，这是一种特别常见的情况:尤其是当动态对象缓慢移动或从静止位置开始移动时(想象一下典型的情况，•汽车在交通灯前临时停车，如图1所示)，异常值的检测极其困难。如果对观察到的场景没有进一步的了解，特别是对于单眼方法，通常不可能区分图像中的静态和移动部分。近年来，基于卷积神经网络的场景理解和语义分割技术得到了很大的发展，利用高层次的推理可以减少图像中静态和动态部分之间的模糊性。考虑到新的卷积架构和模型的开发能够在mobilelembedded gpu上高效、低内存占用的情况下运行，这一点特别有趣。

通过对场景语义的了解，可以检测潜在的动态对象，而不需要显式地跟踪它们。能够对场景中建筑物、车道标志等静态部分进行分割，指导特征提取和匹配。此外，我们提出了一个概率模型，而不是仅仅依赖于逐帧的语义信息。它考虑了所有帧的语义信息，估计出每个映射点的语义。除了语义信息外，我们还使用时间运动信息来讨论某个地图点是动态的还是静态的。当新的观测结果出现时，我们更新地图点的概率参数。为了实现一个实时SLAM系统，我们设计了一种有效的在线概率更新方法，具有较低的内存消耗。在我们的评估中，我们在合成和真实数据集的高度动态情况下显示了更稳定的结果，同时在静态场景中显示了与最先进的方法类似的性

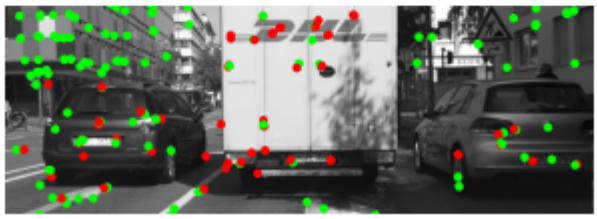
(a) Semantic Segmentation.



(b) Naive approach.



(c) Masking approach.



(d) Our approach.

图1:来自CityScapes数据集的一个例子，该数据集描述了一个困难且高度动态的场景，其中一辆汽车正站在交通灯前。占据了图像的大部分地区属于对象只是暂时静态和将开始慢慢移动,造成基于只对孤立点检测(b)运动的线索的方法失败。在(c)中语义mask忽略所有潜在动态领域的要点,因此无法使用停车辆姿态估计。我们的方法(d)使用深度方差和融合（a）语义信息的点对离群点估计。

能。

II.相关工作

大多数SLAM算法都将动态对象视为异常值。相对于现有的语义SLAM方法，我们建议使用语义信息来选择一组位于静态场景部分的活动特征，以获得更鲁棒的姿态估计，重点是密集的语义三维重建。语义先验是由基于RGB图像的深度模型生成的。

A.动态SLAM

过去，针对视觉SLAM中动态异常点的处理提出了不同的策略。在[4]中，只有经过一定数量的观测后收敛到具有较小方差的深度的活动特征才被用来跟踪。对[5]进行了各种修改，以显式地处理动态对象。在[7]中使用了另一种RANSAC公式，其中调整采样以分布采样点。[9]利用光流在所有特征点的流向图中找到簇，并利用簇将动态对象从静态背景中分割出来。RGB-D相机或立体相机的使用产生了高度可靠和密集的深度地图，在这些情况下，自由空间推理可以用来检测动态对象。如果动态对象移动到以前空闲的区域，并标记为位姿估计[10]的异常值，就会检测到它们。当只有稀疏和有噪声的深度信息可用时，自由空间推理是不可能的。为了处理单目系统中的动态场景，最近的工作集中在多体结构-从运动公式。这里的场景分为多个刚性运动的物体和静态的世界。首先通过运动分割检测目标实例，然后对每个聚类进行帧到帧的转换计算，并使用bundle平差对最终的轨迹[11]进行优化。这里输出的质量取决于运动分割。如果运动较小，分割效果较差，缓慢运动的物体无法被正确检测到。执行时间也远不是实时的。

B.语义SLAM

现有的方法大多结合经典的SLAM和场景的语义分割，利用SLAM系统的位姿图对图像序列进行时间或空间一致性分割。时间或空间一致性可以表示为CRF，在图像[12]、密集体素网格[13]或网格[14]上。由于使用密集的CRF，这些方法中的大多数都不适合在动态场景中进行大规模的实时应用，因为它们的帧率[12]较低。其他方法使用在线更新来实现语义融合[15]，这允许它们实时运行。上述大部分方法都没有将语义信息反馈到姿态估计管道中。在[12]中，利用语义信息对三维模型融合过程中的测量值进行加权。如果多个观察值的语义类不同，则[16]删除点。为了获得致密的三维模型，[12]和[17]采用了立体摄像机。

III.概率语义SLAM

计划SLAM系统建立在ORB SLAM 框架之上[1]，该框架由三个模块(1)跟踪、(2)映射和(3)闭环组成。图2给出了框架的概述。我们提出了一种动态和静态映射点的显式模型，并对其进行了跟踪
我们从前两帧开始基于基本矩阵估计并初始化ORB特征[1]，然后进行全局BA调整，共同优化相机姿态和地图点[11]。为了补偿位姿估计误差，我们使用Lucas-Kanade光流[18]，结合极线约束而不是沿外极线搜索(如[3])来估计直接特征的初始深度。

对每个新的帧ORB特征进行提取，并通过描述子匹配找到对应匹配点。位姿估计用一种基于等速运动模型进行初始化。我们根据描述性和直观性特征对新框架的姿态进行了优化，通过多分辨率多步非线性优化。根据每个金字塔等级的描述和直接特征的数量，我们使用多个回合，我们根据特征点的误差添加或删除特征点。我们只提取关键帧上的新直接特性。如果与当前关键帧没有足够的对应匹配，则创建一个新的关键帧。

利用新帧的估计位姿，可以对三角图进行后期的描述和直接特征提取，从而得到深度估计。在一个固定的关键帧窗口上，通过局部束的调整，对地图点和相机姿态进行联合优化。

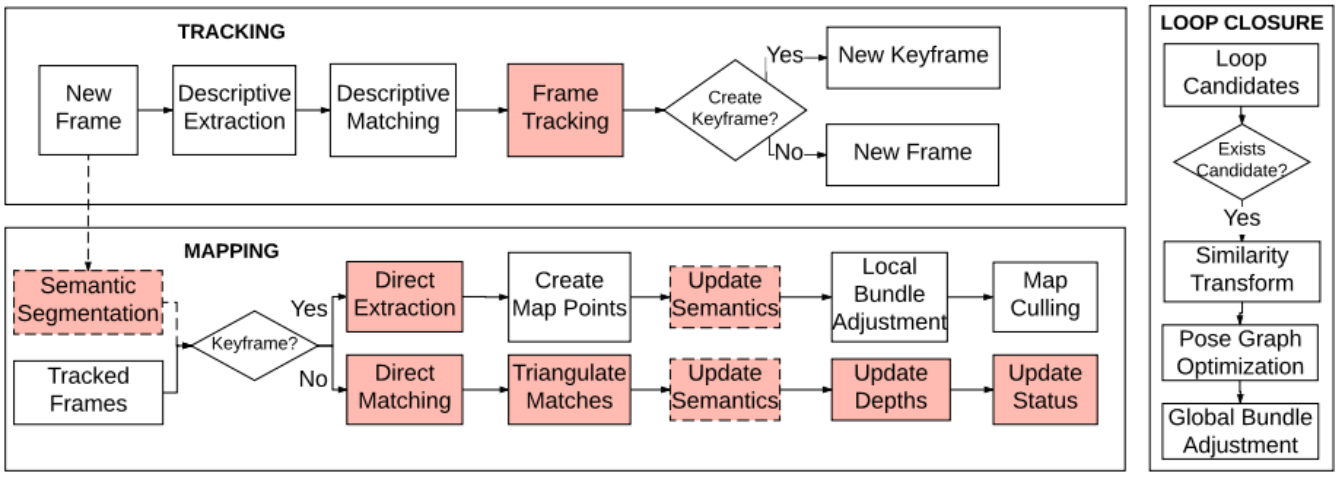


图2:动态SLAM框架概述。将ORB-SLAM方法扩展到直接特征提取和匹配的映射模块中，将其应用于姿态估计的跟踪模型中。我们还引入了一个概率离群模型来更新每个映射点的状态。在增加或修改模块的位姿估计中，只使用活动映射点来集成直接特征，而概率独立点模型用红色表示

A.位姿确定和建图

像ORB这样的描述性特征用于structure-from-motion方法，提供快速可靠的匹配，减少了假阳性对应的数量。地图点通常参数化作为3D点$X\in R^3$并且优化方法是最小重投影误差(见方程2)。此外,描述性特性可以用来识别闭环检测和在一个现有的地图中重定位,可以用于应用程序本地化的自主驾驶和在增强现实应用程序标签。
另一方面，直接特征（直接法）避免了提取关键点和描述符的计算开销。另一方面，为了使优化收敛到全局最小值，需要一个好的初始位姿。这导致需要高帧率或相对较慢的相机运动。基于图像补丁的匹配精度不像描述符比较那么可靠，可能无法检测到假阳性或局部极小值。直接特征对强烈、突然或局部光照变化也很敏感。联

合估计仿射曝光模型[3]只能减少这些影响。然而，直接特征也可以提取在低纹理环境或如果存在强烈的运动模糊。
由于这些互补的特性，我们决定尽可能使用（特征点法）描述性特性。在无法找到足够特性的情况下，我们还使用直接特性（直接法）。
代替通常用于描述特征的3D地图点方法，我们决定使用与直接特征一致的反深度公式。因此，我们可以对这两种特征使用相同的概率模型，从而简化了联合优化中的权重。
我们决定使用ORB特性作为描述性特性，因为它们具有快速提取(快速角+方向)和健壮描述子(BRIEF)。We follow the implementation of [1] to get an equal distribution over the whole image using a grid extraction strategy and a multi-resolution pyramid.(我们使用网格提取策略和多分辨率pvramid实现了[1]在整个图像上的均匀分布)。对于直接特征，我们使用分辨率金字塔和基于网格的特征选择技术提取[37]。
在姿态估计和局部BA优化过程中，将**重投影误差ER和光度误差Ep的加权和减到最小**。
$$E=\eta_R\sum_{M}E_R^T\sum_i^{-1}E_R+\eta_P\sum_NE_P^{T}\sum_i^{-1}E_P\tag{1}$$ 这里η_R,η_P,平衡了重投影和光度误差。因为我们是在逐点推理。不需要对这两个参数进行动态调整。
我们将T_n定义为$SE(3)$将一个表示为世界坐标系的点$X\in R^3$转换为坐标系n的位姿变换，K为摄像机的固有标定矩阵，表示从齐次坐标到笛卡尔坐标的转换，d为关键点的估计深度。
重投影误差由观察到的关键点$(x_i,y_i)^T$的像素距离和匹配的关键点$x^\wedge =(x_i,y_i,1)^T$从坐标系i到坐标系j的投影给出.$$E_R=[x_j,y_j]^T-\pi(KT_j(t^(-1)*i)[d(K^(-1)x^\wedge)])\tag{2}$$ *以同样的方式,光度误差是图像Φ_i上像素(x_i,y_i)周围和它在图像Φ_j上的投影的像素强度差别。$$E_P=\Phi_i(x_i,y_i)-\Phi_j(\pi(KT_j(T_j^{-1}[d(K*{-1}x^\wedge)))\tag{3}$$* 我们接下来为光照变化使用仿射变换模型的。
$$\Phi_i=\frac{1}{t_i}e^{\alpha_i}(I_i-\beta_i)\tag{4}$$ 其中t_i为快门时间，α_i和$beta_i$为每帧估计的仿射变换参数。
采用带鲁棒Huber范数的加权高斯-牛顿法求解非线性最小二乘问题。与[2]相似，我们使用协方差尺度，每一项都用其逆协方差加权来反映每一测量的不确定性。
每次新测量后，通过公式5中的更新进行协方差传播。公式（5）：$$\sum=\frac{\partial E^}{\partial d}\sigma^2(\frac{\partial E^}{\partial d})^T\tag{5}$$

B.概率剔除外点

SLAM的一个核心思想是，通过对每一个重新观测到的地图点的三维位置进行更新，对每一个新的测量值对三维世界的地图进行细化。由于某些度量比其他度量更可靠，因此依靠一种利用度量的方差作为权重的概率方法，可能比天真地求平均值做得更好。

在动态环境中，仅估计地图点的位置是不够的。如果我们使用场景中的所有点执行BA优化，包括动态点，这将导致损坏的优化估计，因为BA优化假定点位置的时间一致性。因此，我们也想知道哪些点对BA优化是足够可靠的。

我们估计了每个地图点的**离群比(inlier ratio)ϕ**，描述了地图点可靠的稳定点的可能性。inlier比值（离群比）可以用多种方法建模，例如，**一些方法保持跟踪成功和不成功的三角剖分[1]的数量。在[19]中，采用概率模型[4]对深度进行联合建模，inlier ratio作为潜在变量**。在这两种情况下，离群比都是通过观察地图点的位置随时间变化来更新的，并根据估计的相机姿态来判断它们是否是动态的。

在单目SLAM中，通过对地图点的运动估计来确定其离群比是很困难的。在缓慢移动的物体的情况下，或者如果一个大的动态物体占据了相机的大部分视角，那么obiect本身就被认为是静态世界的一部分。我们**将语义信息包含在离群比的估计中**，以提供关于地图点是动态的可能性的另一个独立信息源。因此,在深度d和窗比ϕ之外,我们也估计每个地图的语义类c。

当观察到一个地图点时，我们用三角剖分法计算其当前深度估计值d_i，再加上估计的方差τ_i^2。三角剖分法得到的新测量结果的方差，假设关键点在图像中的位置只有用像素精度[19]才能知道。我们还估计匹配

精度$alpha_i\in[0,1]$，如后面所述，并从神经网络中检索关键点的语义类概率$CNN(c_k|I_i)\in[0,1]$。这里$CNN(c_k|I_i)$是网络的输出，可以理解为给定当前图像帧I_i中，一个关键点属于语义类C_k的概率。

我们定义深度测量的似然概率如式6所示。它是基于[19]的，但我们扩展了它的匹配精度。为了简化符号，我们使用$x^-=(1-x).$ $$p(d_i|d,\phi):=\alpha_i[\phi N(d_i|d,\tau_i^2)+\phi^0 U(d_i)]+\alpha_i^-U(d_i) \tag{6}$$ 这个定义背后的直觉是这样的:如果当前键值匹配正确且映射点是静态的。然后两者的匹配精度α;离群比ϕ接近1。因此，假设深度测量d;分布为高斯分布$N(\mu,\sigma^2))$，分布在均值mu附近，方差σ^2。另一方面。如果当前匹配错误，或者测点是动态的，则假设当前深度测量是均匀分布$U(a, b)$且对于平均深度d的估计没有提供任何有用的信息。
与深度的情况类似，我们将映射点的语义建模为网络输出$CNN(c_k|I_i)$和错误匹配的关键点的均匀分布的混合:
$$p(c_k|I_i):=\alpha_iCNN(c_k|I_i)+\alpha^-_iU(C_k) \tag{7}$$ 这允许有效的在线更新和地图点之间的动态和静态平稳过渡.

最后，我们需要定义离群比(inlier ratio）对语义类的依赖关系。结果表明，如果我们将depen lency模型化为Beta分布，就可以得到有效的在线参数更新，如公式8所示。 $$p(\phi|c)=\prod_{k=1}^K(\frac{1}{B(A_k,B_k)}\phi^{A_k-1}(1-\phi)^{B_k-1})^{c_k}\tag{8}$$

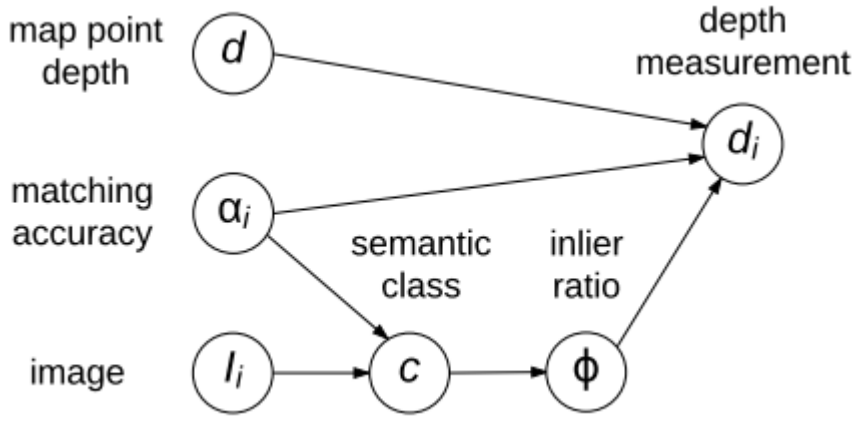


图3:提出的联合概率模型图。显示深度测量d_i、匹配精度α_i和离群比ϕ之间的关系，离群比ϕ取决于CNN从当前帧预测的语义类概率c。 这里的参数c是一个热编码的语义类和$A_k,B_k>0$是固定常数，为每个语义类设置。它们表示某个类是静态或动态的可能性(例如，car类有一个低A_k和高B_k，因为它更有可能是动态的)。相对于深度测量，常量A_k,B_k可以被缩放，从而对语义测量施加或多或少的权重，即相对于运动先验，较高的A_k和B_k更倾向于语义先验. 图3给出了联合模型对深度、离群比和语义类的依赖关系图。测量深度d_i依赖于实际深度d，匹配精度α_i和离群比ϕ，离群比ϕ依赖于语义c。 近似推理导致后验概率结合了三项。第一项包括深度为高斯分布，第二项是基于深度测量的偏最小二乘分布形式的偏最小二乘分布，第三项是对偏最小二乘与语义类的依赖关系进行建模的偏最小二乘分布。
$$p(d,\phi|D,S)=N(d|\mu,\sigma^2)Beta(\phi,a_{obs},b_{obs})Beta(\phi,a_sem,b_sem) \tag{9}$$ 这里$D={d_1,...,d_N}$均为深度测量值，$S={s_1,...,s_N}$均为具有语义信息的观测值，$s_i=(CNN(c_1|I_i),...,CNN(c_K|I_i))$为CNN的输出即为K类的概率密度。

可以看出，所有深度测量值都可以用平均深度μ和深度方差σ^2来概括，相似的，离群率inlier ratio呈beta分布，参数为$a_obs + a_sem$和$b_obs +b_sem$。通过进一步的代数处理，可以得到这些参数的有效在线更新,支持快速更新点的概率模型。

对于没有语义信息的帧，不使用最后一个术语。a_sem和b_sem的语义beta分布参数表示为:
$$a_{sem}=\sum_{k=1}^KA_kp(c_k|S)\tag{10}$$ $$b_{sem}=\sum_{k=1}^KB_kp(c_k|S)\tag{11}$$ 类后验概率$p(c_k|S)$是所有语义测量值的概率融合，见式12。与已有的融合方法[13]相比，式7中的定义根据每个测量α_i的匹配精度，导致加权语义融合. $$p(c_k|S)∝\prod_{i=1}^Ncnn(c_k|I_i)^{\alpha_i}\tag{12}$$ 为估计描述性特征的匹配精度，用汉明距离来比较二元描述符。 $$\alpha^{descriptive}:=1-min(1,\frac{d(f_i,f_j)}

{d_{max}})\tag{13}$$ 对于直接特征，我们使用两个归一化图像块之间的光度差。 $$\alpha^{direct}:=1-min(1,\frac{\Delta \Phi(x_i,x_j)}{\Delta\Phi_{max}}) \tag{14}$$ 在我们的实现中，我们使用反向深度[4]，[20]，来对无穷远处的点进行建模。[20]还表明，反演深度更有可能是高斯分布的。这取决于我们是否使用地图点进行姿态估计(主动)或不主动(非主动)。目前的离群比可计算如式15所示。 $$\phi=\frac{a_{obs}+a_{sem}}{a_{obs}+a_{sem}+b_{obs}+b_{sem}}\tag{15}$$

C.实时语义分割

在高度动态的场景中，图像内容可以快速变化。对于快速移动的相机，我们需要在每一帧中提取新的关键点，以保持足够的活动关键点来进行可靠的跟踪。因此，我们在每个新帧中提取关键点，而不仅仅是在关键帧中。为了得到每个新地图点的一致语义度量，我们在所有新帧上运行语义分割。 我们使用了19类预先训练的[8]模型用于CityScapes数据库。我们按照建议的训练过程将模型细化到其他数据集，使最后一层适应于可用的类集。由于缺少其他类的额外信息(这些信息很容易识别)，仅使用静态和动态类而不是多类标签来训练模型会导致稍微糟糕的结果。