第 58 卷 第 6 期/2021 年 3 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

面向动态场景的语义视觉里程计

卢金，刘宇红，张荣芬*

贵州大学大数据与信息工程学院，贵州 贵阳 550025

摘要 针对传统视觉同时定位与地图构建(vSLAM)相机跟踪模块在动态环境中无法精确定位的问题，提出一种基于语义的视觉里程计。首先，在利用金字塔 Lucas-Kanade 光流追踪匹配帧间特征点的同时，对图像进行像素级的语义分割。然后，将语义信息与几何特征紧密结合用以准确地剔除图像中的外点，使得位姿估计和建图仅依靠图像中值得信赖的静态特征点。最后，提出了一种多尺度的随机抽样一致(RANSAC)方案，对匹配点进行步进采样，每步使用不同的尺度因子，在降低外点检测时间的同时，提高了外点检测的鲁棒性。在 TUM 数据集上的实验结果表明，在高动态序列中，相比于 ORB-SLAM2，本文方案的绝对轨迹误差和相对位姿误差改善了 90%以上，而相比于同类型的 DS-SLAM，本文方案在降低外点检测时间 30%～40%的情况下，提升了位姿估计的鲁棒性。

关键词 成像系统；运动估计和光流；视觉里程计；动态场景；多尺度随机抽样一致；语义分割

中图分类号 TP391；TP18 **文献标志码** A **doi**：10.3788/LOP202158.0611001

Semantic-Based Visual Odometry Towards Dynamic Scenes

Lu Jin, Liu Yuhong, Zhang Rongfen*

College of Big Data and Information Engineering, Guizhou University, Guiyang, Guizhou 550025, *China*

Abstract To deal with the problem that the camera tracking module of traditional visual simultaneous localization and mapping (vSLAM) can′t make pose estimation accurately, a semantic-based visual odometry is proposed. First, while using pyramid Lucas-Kanade optical flow to track and match the inter-frame feature points, the frame is pixel-wisely segmented. Then, the semantic information and geometric features are combined closely to accurately remove the outliers in the frame, thus the pose estimation and mapping can rely only on the trusted static feature points in the frame. Finally, a multi-scale random sample consensus (RANSAC) scheme is proposed. The matching points are sampled step by step, and different scale threshold are used for each step, which can reduce the detection time and improve the robustness of outliers simultaneously. Experimental results on the TUM data set show that, compared with ORB-SLAM2, the absolute trajectory error and relative pose error of the proposed system are improved by more than 90% in the high dynamic sequence. And the proposed scheme reduced the detection time by 30%－40% while the robustness of pose estimation is improved when compared with similar DS-SLAM.

Key words imaging systems; motion estimation and optical flow; visual odometry; dynamic scene; multi-scale random sample consensus; semantic segmentation

OCIS codes 110.4153; 200.4260; 100.3010

1 引 言

作为机器人乃至计算机视觉领域的重要技术之一，视觉同时定位与地图构建(vSLAM)在过去几十年吸引了全世界大量研究者的注意，一批经典的 vSLAM 框架，如 PTAM[1]、ORB-SLAM2[2]、LSD-

收稿日期：2020-07-10；**修回日期**：2020-07-30；**录用日期**：2020-08-27

基金项目：贵州省科技计划项目(黔科合基础[2019]1099 号)

* **E-mail**：rfzhang@gzu.edu.cn

SLAM[3]、SVO[4]、DSO[5]等因此产生。其中，相机跟踪是 vSLAM 的核心，正确的相机跟踪能为搭载该技术的机器人提供理解环境的必要信息，但受限于其对环境中物体均为静态刚体的假设[6]，在处理动态的复杂场景时，经典框架由于不能获得正确的帧间匹配，导致其定位与建图精度大大降低。解决上述问题的思路是在里程计部分准确剔除图像中的外点，即动态目标引入的动点，使机器人仅依靠值得信赖的静态点来进行相机跟踪。

Kundu 等[7]通过计算基本矩阵来判定特征点的动静状态，在对极几何约束中，当前帧与上一帧的匹配点应该位于基本矩阵对应的极线附近，如果某点与极线相去甚远，则该点很可能是动点。该方法的关键是能相对可靠地估计出基本矩阵，一旦计算出相对准确的基本矩阵，就能较为精确地检测并剔除图像中的外点。为了避开基本矩阵的计算，Migliore 等[8]通过三角观测原理来判定点的运动状态，他们在一种概率滤波框架下连续在三个视角观测三条投影线的交点，如果物体发生了运动，则这三条线的交点位置将发生改变甚至不相交，类似地，文献[9]通过德洛奈三角来对特征点进行动静分割。这些方案能够通过一定的几何约束来计算出比较明显的外点，然而由于物体表面的纹理、亮度等多种因素，几何法并不能完整地分割出运动物体的整个轮廓。

为了获取外点所属物体的轮廓，Klappstein 等[10]通过计算特征点违背光流法灰度不变假设的程度，来判定该点是否是动点，然后通过图割理论来粗略分割出该点所属物体的轮廓，进而剔除轮廓内的特征点。林志林等[11]在视觉里程计中引入场景流计算模型，并构造图像特征的高斯混合模型进行运动物体检测，同时按照场景中物体的运动模型构造虚拟地图点，最终结合高斯混合模型和虚拟地图点进一步筛选运动物体，该方案在运动物体占据图像大部分区域时，仍能获取足够的匹配点来保证相机位姿估计的正常进行。

利用深度学习对图像进行像素级分割，可以弥补几何法不易得到运动物体的整个轮廓的缺陷。Xiao 等[12]使用单独的线程运行 SSD 来获取动态物体的语义信息，在跟踪线程中通过一种选择性跟踪算法对动态对象的特征进行处理，可以显著降低姿态估计的误差。文献[13]基于 YOLO v3 提出了一种适用于动态场景的高效 SLAM，首先利用 YOLO 网络进行目标检测，然后利用基于深度图的漫溢填充算法来精确地分割出检测目标的轮廓，最后构建不包含动态物体的静态语义地图。仅仅利用语义信息存在的一个缺陷是无法判定物体的运动状态，只能按照经验从图像中去除运动概率较大的物体，这种朴素的去除方法导致许多静止的物体也从图像中剔除，使得相机无法获取足够的可靠特征点用于位姿估计。

近年来，许多研究都尝试将语义信息与几何约束相结合，以获得更好的外点去除结果。Brasch 等[14]将特征法、直接法以及语义信息融合为一个概率模型，用于估计单目相机视野中物体的运动概率，当物体的运动概率超过一定阈值时判定其为动态物，并将该物体上的特征点剔除，该方法的优势是在动态物占据相机大部分视野时仍能做出较好估计。Yu 等[15]提出一种 DS-SLAM 方案，将 SegNet 与光流法相结合，在利用光流法对 RGB 帧图进行运动一致性检测的同时，利用 SegNet 对其进行语义分割，继而判定外点是否位于人体的分割区域内，如果是，则将位于该区域内的所有特征点进行剔除。类似的工作还有文献[16]首先对输入图像进行实例分割，并采用文献[17]中的边界检测方法来调整语义分割的边界，进一步提升分割精度，使得外点剔除更加完整。这类方案相较于前两种能取得更为显著的外点剔除效果，但往往需要在精度和实时性中做出取舍。

本文以 ORB-SLAM2 里程计部分为基础，一方面选用在实时性和精度方面有较好平衡的 SegNet[18]作为语义分割网络获取语义信息，另一方面利用金字塔 LK(Lucas-Kanade)光流法直接追踪特征点以取代相对复杂的 ORB(Oriented Fast and Rotated Brief)描述子的提取与匹配，并将语义信息与几何信息有效结合用于准确剔除图像中的外点。其中，本文提出了一种多尺度的随机抽样一致(RANSAC)方案，该方案在实现更为精确的基本矩阵计算的同时，保留了更多内点以及减少了动点的检测时间。

2 算法研究

本文提出的视觉里程计如图 1 虚线框所示，该方案在 ORB-SLAM2 相机追踪部分的基础上进行优化，新增了语义分割线程用以获取语义信息，同时在追踪线程加入动态目标外点检测和移除环节。其中，在外点移除环节，系统将语义信息与几何信息有效结合用于精确剔除外点。对于相机输入的每一帧 RGB 图像，同时将其送入语义分割线程和追踪线程。在追踪线程，提取ORB特征点后，首先采用金

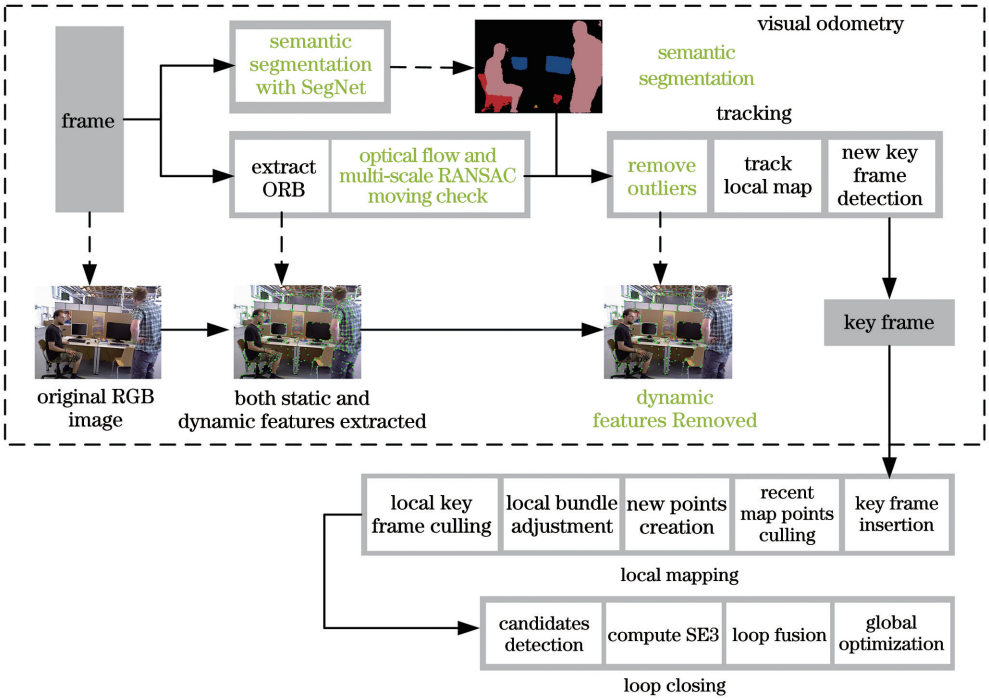

图 1　本文视觉里程计的系统结构

Fig. 1　System structure of proposed visual odometry

字塔 LK 光流法实现前后两帧特征点的快速匹配，然后利用基于多尺度的 RANSAC 选取可靠匹配对，并利用这些匹配计算基础矩阵 $\boldsymbol{F}$，随后按照计算出的 $\boldsymbol{F}$ 利用对极几何对所有匹配对进行运动一致性验证，来检测出极大部分的外点，继而进一步结合语义分割线程获取的语义信息来剔除动态物体上的动态特征点，系统最终仅利用图像中的静态特征点来估计位姿。

2.1　语义分割

本文在语义分割线程中采用了基于 Caffe 的 SegNet 作为语义分割网络，SegNet 是由剑桥大学计算机视觉和机器人小组开发的多类别语义分割网络，该网络基于深度的编码解码架构，能对图像进行像素级分割，其网络架构如图 2 所示，其中与编码网络一一对应的层次化解码网络是 SegNet 的核心。SegNet 的编码网络与 VGG16[19] 网络相似，但去掉了全连接层，这使得 SegNet 的编码网络相较于其他架构的网络，在规模更小的同时更易于端对端的训练。同时，由于 SegNet 最初的设计目的是对室外的道路场景进行分割，因此该网络对场景外观、形状、空间关系等方面都能较好处理，对于边界信息的处理也更为精确。

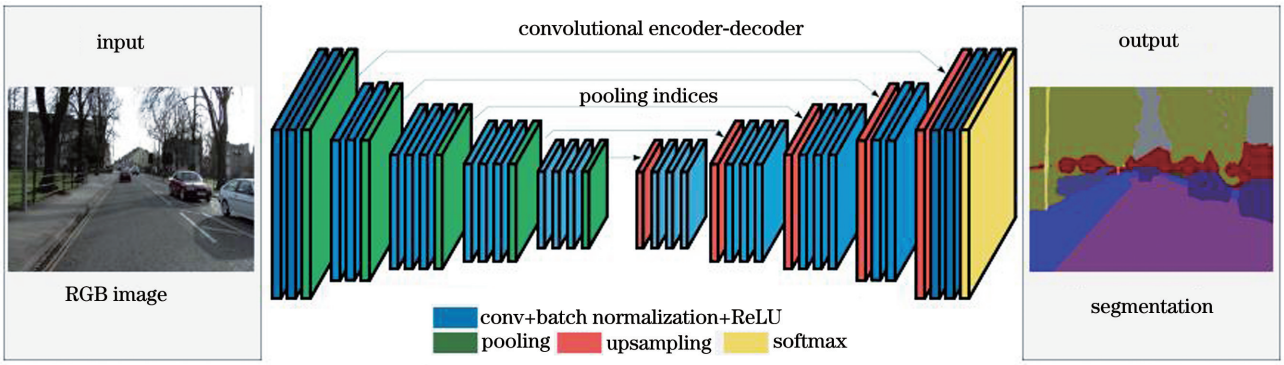

图 2　SegNet 网络架构

Fig. 2　Network architecture of SegNet

为了对室内场景进行分割，本文在 PASCAL VOC 数据集上对 SegNet 进行训练，该数据集包括人、猫、狗等室内常见的动态物体。图 3 为网络对室内场景的分割效果，为了适应 PASCAL 数据集中图

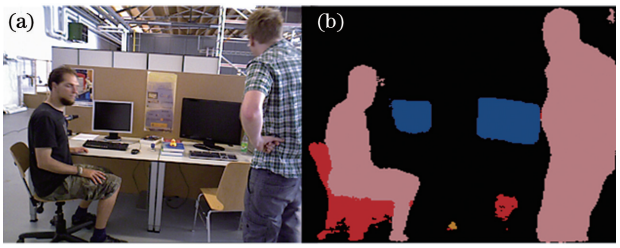

图 3　语义分割效果。(a) 原图;(b) 分割结果

Fig. 3　Results of semantic segmentation. (a) Original image; (b) segmentation result

片的不同尺寸,将网络的输入尺寸调整为 224×224,尽管分辨率降低使得分割精度有所降低,但网络仍能较为精确地分割出图像中人体的轮廓,用于本文的外点剔除工作。

2.2　运动一致性检测

2.2.1　多尺度 RANSAC

为了避开繁复的描述子计算与匹配,本文采用金字塔 LK 光流法来跟踪匹配前后两帧图像间的特征点。金字塔 LK 光流基于特征点邻域内像素块的灰度不变假设来估计像素在图像间的运动,由于边缘特征点的邻域不一定都位于图像上,同时由于图像中物体边缘特征点的差异性极小,导致金字塔 LK 光流对图像边缘特征点以及物体边缘特征点的跟踪鲁棒性较差[20]。因而在计算基础矩阵前,一般需要采用 RANSAC 先对匹配对进行降噪,以剔除那些不值得信赖的匹配对。标准的 RANSAC 流程如下:

1) 在特征点集中选取最小匹配对(如 4 对)用于估计基础矩阵 $\boldsymbol{F}$;

2) 在当前帧中,计算其他特征点与由 $\boldsymbol{F}$ 所确定的极线 L 的距离 D;

3) 按照规定的阈值 e,计算出所有 $D<e$ 的特征点,即内点集;

4) 重复步骤 1~3 直至达到迭代次数 N,选取所有迭代中内点最多的一次作为最终方案。

标准 RANSAC 的一个弊端是,当动态物占据相机的大部分视野时,如果此时直接对匹配对进行采样,则其鲁棒性将会大大降低[14]。受到文献[15]和[21]的启发,本文在 RANSAC 之前,先将位于图像边缘的匹配对排除,然后对其他每个匹配对,以匹配对为中心,选取尺寸为 3×3 的窗口来计算灰度残差,如果对应窗口的灰度残差过大,则排除该窗口中心的匹配对。同时,对 RANSAC 策略进行改进,通过设置两次稍大尺度的阈值步进执行 RANSAC,来替代一次极小阈值的 RANSAC。最终利用最后一次 RANSAC 获取的匹配点来计算基本矩阵,并按照该矩阵确定的极线来判定特征点的运动状态。

RANSAC 阈值尺度设置对经验依赖较多,阈值过大会导致图像中留下更多的外点,阈值过小会导致过多的静态点被剔除,且由于 RANSAC 本身的随机性,更小的尺度不一定会带来更好的结果。为了确定每一步 RANSAC 的尺度因子,本文对多组不同尺度的 RANSAC 策略进行了对比考察。图 4 为同一场景在不同 RANSAC 下的外点剔除效果,场景中两人都有不同程度的运动,因而分布于两人身上的特征点都为外点。其中图 4(a)~(c)为采用本文的多尺度策略,图 4(d)为采用 DS-SLAM 中使用的尺度为 0.1 的标准 RANSAC。由图 4 可知,尽管图 4(c)采用了比图 4(b)更低的尺度,但其外点剔除效果反而更差,而且该方案获取的内点数量已经低于标准 RANSAC[图 4(d)],在高动态环境中,这极易导致相机追踪失败。在四种方案中,图 4(b)中遗留在人体身上的特征点最少,且相比于图 4(d),保留了更多的有效内点。因而,本文将两次阈值尺度分别设置为 1 和 0.2。事实上,相比于标准 RANSAC,多尺度 RANSAC 能够剔除绝大部分的外点,得到更好的外点剔除效果;同时,由于阈值的提高,降低了计算的复杂度,因而虽然执行了多步,但其耗时反而较标准策略更低,这点在 3.3 节中,本

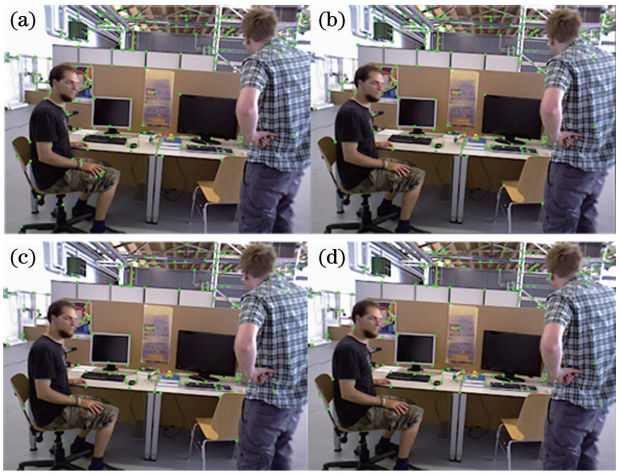

图 4　不同尺度 RANSAC 的效果对比。(a)尺度为 1 和 0.3,内点数量为 238;(b)尺度为 1 和 0.2,内点数量为 210;(c)尺度为 1 和 0.1,内点数量为 159;(d)尺度为 0.1,内点数量为 168

Fig. 4　RANSAC effect comparison under different scales. (a) Scale is 1 and 0.3, number of inliers is 238; (b) scale is 1 and 0.2, number of inliers is 210; (c) scale is 1 and 0.1, number of inliers is 159; (d) scale is 0.1, number of inliers is 168

文系统与 DS-SLAM 在动点检测环节的用时对比中也有所体现。

2.2.2　对极几何约束

运用上述多尺度 RANSAC 选取出可靠匹配对，并计算出基础矩阵 $\boldsymbol{F}$ 后，采用如图 5 所示的对极几何对所有匹配对进行运动一致性验证，以判定图中特征点的运动状态。图中，I_1、I_2 分别代表上一帧与当前帧的成像平面，C_1、C_2 为对应的相机光心，p_1、p_2 为 I_1、I_2 中的一对匹配特征点，则 C_1 与 C_2 的连线称为基线，由基线和特征点 p_1 组成的平面与相机平面的交线 L_1、L_2 称为极线，交点 e_1、e_2 称作极点。理想情况下，如果 p_1、p_2 正确匹配且它们对应的空间点是静止的话，则由光心和特征点组成的两条射线会相交于一点 P，P 即为对应的空间点，此时 p_1、p_2 的归一化平面坐标 $\boldsymbol{p}_1=(u_1,v_1,1)$，$\boldsymbol{p}_2=(u_2,v_2,1)$ 与基本矩阵 $\boldsymbol{F}$ 满足

$$\boldsymbol{p}_2\boldsymbol{F}\boldsymbol{p}_1=0。\tag{1}$$

事实上，受误匹配和外点等因素的影响，$\boldsymbol{p}_2$ 往往不严格位于极线 L_2 上，如图 5(b)所示，此时计算 $\boldsymbol{p}_2$ 到 L_2 的距离，

$$D=\frac{|\boldsymbol{p}_2^{\mathrm{T}}\boldsymbol{F}\boldsymbol{p}_1|}{\sqrt{\|X\|^2+\|Y\|^2}}。\tag{2}$$

当距离大于设定的阈值时，即认为该点为外点。其中 X、Y 为极线 L_2 在三维坐标中向量化后对应的坐标值，该坐标值表示为

$$\boldsymbol{L}_2=\begin{bmatrix}X\\Y\\Z\end{bmatrix}=\boldsymbol{F}\boldsymbol{p}_1=\boldsymbol{F}\begin{bmatrix}u_1\\v_1\\1\end{bmatrix}。\tag{3}$$

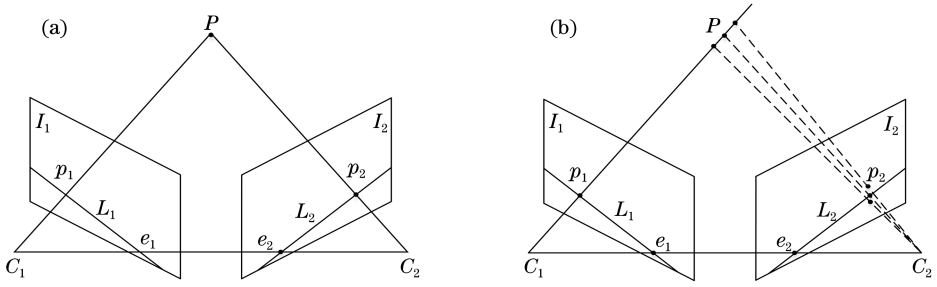

图 5　对极几何约束。(a)p_2 位于极线 L_2 上；(b)p_2 不严格位于极线 L_2 上

Fig. 5　Epipolar geometric constraints. (a) p_2 is on the polar line L_2; (b) p_2 is not strictly on the polar line L_2

2.3　外点剔除

在分别获取到当前帧的外点集 O 以及语义分割信息后，即可进行动态物体的判定及最终的外点剔除。本文的外点剔除策略如算法 1 所示，如果语义分割信息中存在如人等极有可能发生运动的物体的分割区域 M，且 O 中存在点 o 与 M 中的点 m 处在图像的同一坐标位置，则判定该物体为动态物，随即将该物体分割区域从关键点 K 中剔除。由图 6(a)可知，在 ORB 特征提取阶段，系统对所有的包括动态物体上的特征点都进行了提取；经过动态一致性检测并最终结合语义信息将外点剔除后，图像中仅剩下了静态特征点如图 6(b)所示，系统后续便仅依靠这些静态点来进行位姿估计和建图。

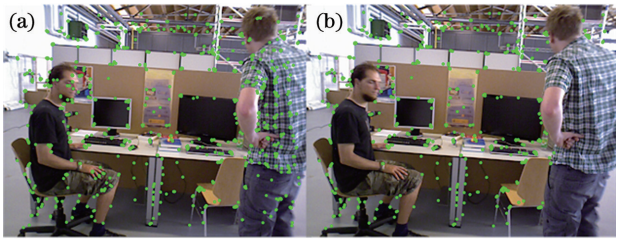

图 6　外点剔除。(a)提取了所有的特征点；(b)剔除了位于人体身上的外点

Fig. 6　Outliers removing. (a) All feature points are extracted; (b) outliers lie on people are removed

Algorithm 1 dynamic points removing algorithm

Input: Dynamic points, O; Semantic mask of most likely moving objects, M; KeyPoints, K;

Output: The set of inliers, I;

```
if M not empty then
for  point o, m in O, M do
    if o=m then
        remove M from K
        leave the loop
    end if
end for
end if
```

3　实　　验

3.1　实验数据集及环境

为了验证本文方案的有效性，选用 TUM RGB-D[22] 数据集中的 8 个动态帧序列来对系统进行测试，并在 3.2 节和 3.3 节分别和原始的 ORB-

SLAM2 以及与本文方案类似的 DS-SLAM 在里程计部分进行了对比。实验环境为 Ubuntu 16.04、Intel i7-9700f CPU、NVIDIA GTX2060 GPU，系统配备 16 G 内存。

TUM 数据集由德国慕尼黑工业大学采集发布，该数据集提供了精确的相机实际运动轨迹以及完备的评估方案。TUM 提供了 8 个序列专门用于评估系统在高、低动态环境下的表现，这 8 个序列由 4 个低动态帧序列和 4 个高动态帧序列组成。其中，4 个低动态序列分别为 freiburg3_sitting_static、freiburg3_sitting_halfsphere 、freiburg3_sitting_rpy、freiburg3_sitting_xyz；4 个高动态序列分别为 freiburg3_walking_static、freiburg3_walking_halfsphere、freiburg3_walking_rpy、freiburg3_walking_xyz。按照 TUM 的命名方式，第一个下划线前的字符表示相机的内参代号 、第二个下划线之前的字符表示图像中人物的运动状态、最后一个字符表示相机的运动轨迹，如 freiburg3_walking_xyz 表示拍摄该序列所用相机的内参代号为 freiburg3，图像中人物的运动状态为走来走去，同时相机沿着 xyz 三个轴进行了运动。为了简化表示，本文后续分别用 sS、sH、sR、sX、wS、wH、wR、wX 来表示上述序列。

评价指标包含：1）绝对路径轨迹误差（ATE），该指标表述相机轨迹的全局一致性；2）相对位姿误差（RPE），该指标表述相机的平移误差以及相对旋转误差。其中，每个指标下又对 4 个参数进行统计，分别是均方根误差（RMSE）、平均误差（Mean）、中值误差（Median）以及标准偏差（S. D.），一般认为 RMSE 和 S. D. 更能体现系统的鲁棒性和稳定性。

3.2 与 ORB-SLAM2 的对比

本文方案基于 ORB-SLAM2 里程计部分调整而来，作为目前最杰出、稳定的 SLAM 系统之一，ORB-SLAM2 在静态环境中有着出色的表现。表 1～3 分别为本文方案与原始 ORB-SLAM2 在 ATE、RPE 两个指标上的定量对比。对比各表中数

表 1 绝对路径轨迹误差典型值

Table 1 Typical value of ATE

Sequency	ORB-SLAM2 /m				Proposed /m				Improvement /%			
	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.
wX	0.565505	0.528696	0.515921	0.200691	0.019148	0.015891	0.013613	0.010682	96.61	96.99	97.36	94.68
wH	0.327989	0.275986	0.232572	0.177225	0.028845	0.024179	0.020507	0.015730	91.21	91.24	91.18	91.12
wR	0.817879	0.695593	0.642242	0.430206	0.407781	0.352362	0.270319	0.205247	50.14	49.34	57.91	52.29
wS	0.409268	0.369913	0.293660	0.175114	0.007302	0.006431	0.006031	0.003459	98.22	98.26	97.95	98.02
sX	0.009275	0.007939	0.007251	0.004796	0.009962	0.008540	0.007845	0.005129	−7.41	−7.57	−8.19	−6.94
sH	0.027882	0.024288	0.022784	0.013692	0.014589	0.012853	0.011611	0.006902	47.68	47.08	49.04	49.59
sR	0.021513	0.016177	0.011756	0.014181	0.016531	0.012956	0.009905	0.010268	23.16	19.91	15.75	27.59
sS	0.007698	0.006775	0.006045	0.003655	0.006142	0.005233	0.004683	0.003216	20.21	22.76	22.53	12.01

表 2 相对平移误差典型值

Table 2 Typical value of relative translation error

Sequency	ORB-SLAM2 /m				Proposed /m				Improvement /%			
	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.
wX	0.825981	0.692976	0.648657	0.449478	0.028086	0.023557	0.020073	0.015293	96.60	96.60	96.91	96.60
wH	0.502363	0.403237	0.436454	0.299615	0.040362	0.035091	0.032131	0.019942	91.97	91.30	92.64	93.34
wR	1.212279	1.006164	0.943469	0.676205	0.138982	0.087917	0.042314	0.102784	88.54	91.26	95.52	84.80
wS	0.585281	0.403727	0.157596	0.423743	0.010569	0.009465	0.008933	0.004703	98.19	97.66	94.33	98.89
sX	0.013602	0.011845	0.010865	0.006688	0.014602	0.012807	0.011729	0.007015	−7.35	−8.12	−7.95	−4.89
sH	0.040732	0.033476	0.028965	0.023205	0.020813	0.018533	0.016983	0.009471	48.90	44.64	41.37	59.19
sR	0.030898	0.025071	0.020812	0.018059	0.024480	0.020617	0.017343	0.013200	20.77	17.77	16.67	26.91
sS	0.012007	0.010637	0.009772	0.005570	0.009133	0.008002	0.007147	0.004403	23.94	24.77	26.86	20.95

表 3　相对旋转误差典型值

Table 3　Typical value of relative rotation error

Sequency	ORB-SLAM2 /(°)				Proposed /(°)				Improvement /%			
	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.
wX	14.812930	12.411189	11.075135	8.086117	0.717936	0.554291	0.449855	0.456282	95.15	95.53	95.94	94.36
wH	13.379170	11.226538	14.220488	7.277847	0.930220	0.806664	0.727123	0.463252	93.05	92.81	94.89	93.63
wR	22.021472	17.877791	16.175450	12.858064	2.780387	1.798397	0.922173	2.013343	87.37	89.94	94.30	84.34
wS	10.334787	7.085263	1.830560	7.523754	0.286392	0.257924	0.242004	0.124482	97.23	96.36	86.78	98.35
sX	0.578052	0.494635	0.422254	0.299133	0.590790	0.509834	0.442760	0.298500	−2.20	−3.07	−4.86	0.21
sH	1.030726	0.924055	0.857998	0.456638	0.716727	0.644671	0.600926	0.313205	30.46	30.23	29.96	31.41
sR	0.882169	0.767921	0.700110	0.434188	0.755252	0.670573	0.623672	0.347473	14.39	12.68	10.92	19.97
sS	0.336292	0.303505	0.286266	0.144834	0.316251	0.283073	0.264374	0.141012	5.96	6.73	7.65	2.64

据可以看出，得益于语义信息的加入，本文方案在高动态帧序列中提升明显，其中在 wX、wH、wS 三个序列下，各个指标的精度均提升了 90% 以上；尽管 ORB-SLAM2 在低动态环境中表现优异，但本文在除 sX 外的其余三个低动态序列下仍有不同程度的提升，其中在 RMSE 方面有着最低 20%、最大 47% 的提升，在 S. D 方面有着最低 12%、最高 49% 的提升。

为了直观展示两个系统在高动态环境中的性能差异，将 wX 序列下两个系统的绝对运动轨迹误差和相对平移误差分别可视化为图 7、8，并进行对比。从图中可以看出，ORB-SLAM2 偏离相机实际路径较为严重，而本文方法估计的路径与相机实际路径大致相同，同时在大多数帧列中，ORB-SLAM2 的相对位移最大误差超过了 0.7，而本文方案在保持大部分帧列相对位移误差小于 0.03 的情况下，最大误差低于 0.12。

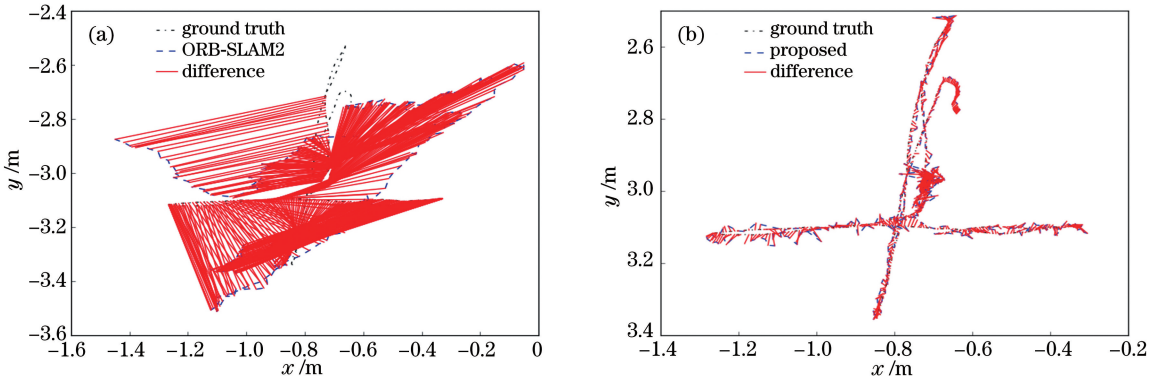

图 7　绝对路径轨迹误差对比。(a)ORB-SLAM2；(b)本文方案

Fig. 7　Comparison of ATE. (a) ORB-SLAM2; (b) proposed scheme

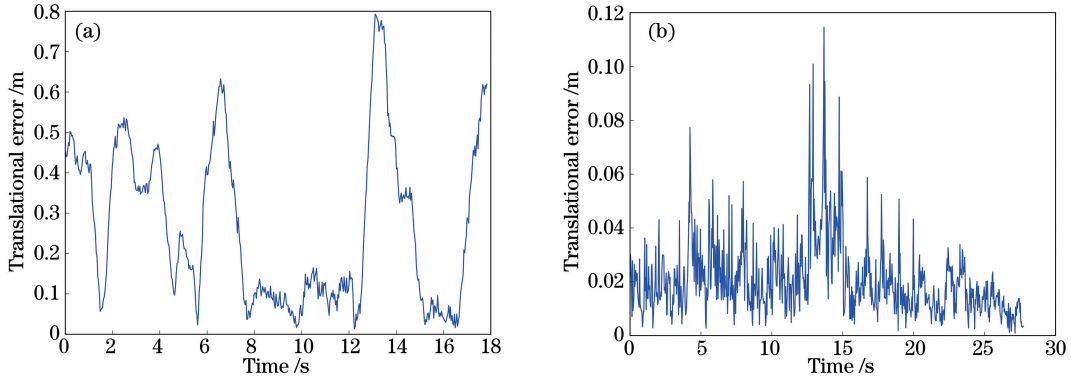

图 8　相对平移误差对比。(a)ORB-SLAM；(b)本文方案

Fig. 8　Comparison of relative translation error. (a) ORB-SLAM2; (b) proposed scheme

3.3 与 DS-SLAM 的对比

为了进一步验证本文方案的有效性，在同样的实验环境中，对本文系统与文献[15]提出的类似系统 DS-SLAM 在相机跟踪部分进行了定量对比。DS-SLAM 在所有面向动态环境的 SLAM 中表现优异，且同样以 ORB-SLAM2 为基础，以 SegNet 为分割网络。与本文不同的是，DS-SLAM 在运动一致性检测阶段采用的是标准的 RANSAC 方案。表 4～6 分别给出了本文系统与 DS-SLAM 在各个序列下的表现对比，表7为两个方案在运动一致性

表 4 绝对运动轨迹误差典型值

Table 4 Typical value of ATE

Sequency	DS-SLAM /m				Proposed /m				Improvement /%			
	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.
wX	0.022180	0.016869	0.013272	0.014402	0.019148	0.015891	0.013613	0.010682	13.67	5.80	−2.57	25.83
wH	0.032083	0.026748	0.022648	0.017715	0.028845	0.024179	0.020507	0.015730	10.09	9.60	9.45	11.21
wR	0.433820	0.368918	0.249150	0.228252	0.407781	0.352362	0.270319	0.205247	6.00	4.49	−8.50	10.08
wS	0.007709	0.006979	0.006576	0.003275	0.007302	0.006431	0.006031	0.003459	5.28	7.85	8.29	−5.62
sX	0.010339	0.008831	0.007981	0.005377	0.009962	0.008540	0.007845	0.005129	3.65	3.30	1.70	4.61
sH	0.014816	0.013229	0.011732	0.006672	0.014589	0.012853	0.011611	0.006902	1.53	2.84	1.03	−3.45
sR	0.020242	0.015779	0.011601	0.012680	0.016531	0.012956	0.009905	0.010268	18.33	17.89	14.62	19.02
sS	0.006142	0.005233	0.004683	0.003216	0.006273	0.005461	0.004728	0.003085	−2.13	−4.36	−0.96	4.07

表 5 相对平移误差典型值

Table 5 Typical value of translation

Sequency	DS-SLAM /m				Proposed /m				Improvement /%			
	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.
wX	0.032488	0.025585	0.020938	0.020022	0.028086	0.023557	0.020073	0.015293	13.55	7.93	4.13	23.62
wH	0.045461	0.039412	0.035235	0.022658	0.040362	0.035091	0.032131	0.019942	11.22	10.96	8.81	11.99
wR	0.148749	0.094158	0.045830	0.112832	0.138982	0.087917	0.042314	0.102784	6.57	6.63	7.67	8.91
wS	0.010977	0.009989	0.009499	0.004552	0.010569	0.009465	0.008933	0.004703	3.72	5.25	5.96	−3.32
sX	0.014969	0.013095	0.012071	0.007252	0.014602	0.012807	0.011729	0.007015	2.45	2.20	2.83	3.27
sH	0.021379	0.019180	0.017768	0.009444	0.020813	0.018533	0.016983	0.009471	2.65	3.37	4.42	−0.29
sR	0.028873	0.024268	0.020204	0.015643	0.024480	0.020617	0.017343	0.013200	15.21	15.04	14.16	15.62
sS	0.009217	0.008144	0.007363	0.004316	0.009133	0.008002	0.007147	0.004403	0.91	1.74	2.93	−2.02

表 6 相对旋转误差典型值

Table 6 Typical value of rotation

Sequency	DS-SLAM /(°)				Proposed /(°)				Improvement /%			
	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.	RMSE	Mean	Median	S. D.
wX	0.768973	0.583035	0.460229	0.501387	0.717936	0.554291	0.449855	0.456282	6.64	4.93	2.25	9.00
wH	0.983289	0.862411	0.769254	0.472340	0.930220	0.806664	0.727123	0.463252	5.40	6.46	5.48	1.92
wR	3.013413	1.909141	0.996452	2.320810	2.780387	1.798397	0.922173	2.013343	7.73	5.80	7.45	13.25
wS	0.285163	0.261389	0.250628	0.113990	0.286392	0.257924	0.242004	0.124482	−0.43	1.33	3.44	−9.20
sX	0.577467	0.493059	0.420298	0.300600	0.590790	0.509834	0.442760	0.298500	−2.31	−3.40	−5.34	0.70
sH	0.778858	0.698167	0.649726	0.345227	0.716727	0.644671	0.600926	0.313205	7.98	7.66	7.51	9.28
sR	0.863546	0.760713	0.701008	0.408689	0.755252	0.670573	0.623672	0.347473	12.54	11.85	11.03	14.98
sS	0.308551	0.276659	0.259871	0.136614	0.316251	0.283073	0.264374	0.141012	−2.50	−2.32	−1.73	−3.22

表 7 运动一致性检测时间消耗

Table 7 Time consuming of moving consistency check

Sequency	DS-SLAM /ms	Proposed /ms	Reduced /%
wX	0.019487	0.013030	33.14
wH	0.018344	0.010929	40.42
wR	0.017183	0.010957	36.23
wS	0.016712	0.015695	6.09
sX	0.017424	0.014163	18.72
sH	0.018717	0.013182	29.58
sR	0.016412	0.011832	27.90
sS	0.014110	0.013517	4.20

检测方面的耗时对比。

由表 5～7 可知，在运动幅度稍低的 wS 和低运动的 sX、sH、sS 等几个序列中，本文方案虽然提升不明显，但将外点检测时间最大减少了 29%；在 wX、wH、wR 等高动态序列中，本文方案在降低外点检测时间 33%～40%的同时，将绝对运动轨迹、相对平移、相对旋转等指标在 RMSE 方面分别提升了 6%～13.6%、6.5%～13.5%、5.4%～7.7%，而在 S.D 方面，也有最大 1/4 的改善。此外，在低动态序列 sR 中，由于相机产生的运动几乎为纯旋转运动，这给相机的位姿估计带来难度，但得益于本文多尺度 RANSAC 带来的更为稳健的动点剔除结果，sR 序列下本文系统在各个指标上均有 15%左右的提升。综上可见，本文系统相较于 DS-SLAM 实时性更高，且具有更好的鲁棒性。

4 结 论

本文提出了一种面向动态环境的语义视觉里程计方案。在对图像进行 ORB 特征提取的同时，对其进行语义分割，获取高层次语义信息，并在追踪线程将图像几何信息与语义信息结合用于剔除图像中的动态点，最终系统仅靠图像中值得信赖的静态点来进行位姿估计。其中，语义分割网络选取了在实时性和精度方面有较好平衡的 SegNet 网络；在外点检测阶段，使用金字塔 LK 光流法来跟踪匹配特征点以取代繁复的特征点描述子的计算与匹配；同时，对标准 RANSAC 策略进行了改进，提出了一种多尺度的 RANSAC 方案，该方案在降低外点检测时间的同时增加了外点检测的鲁棒性。在 TUM RGB-D 的动态序列下的实验结果表明，在室内高动态环境中，本文系统在精度和鲁棒性方面远优于 ORB-SLAM2，各项指标提升均超过 90%，同时在与同类型的 DS-SLAM 的量化对比中，本文提出的多尺度 RANSAC 在降低外点检测时间多达 40%的同时，提升系统最大 1/4 的精度。然而，本文系统仍然存在可以优化的环节，其中，语义分割网络在分割精度和实时性方面仍值得提升，后续将适配性能和实时性更好的网络；此外，本文专注于定位精度和鲁棒性的提升，接下来将在语义建图方面进行改善，从而使搭载该系统的机器人可以执行更高层次的任务。

参 考 文 献

[1] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C]//2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, November 13-16, 2007, Nara, Japan. New York: IEEE Press, 2007: 225-234.

[2] Mur-Artal R, Tardós J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[3] Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM[M]. Cham: Springer International Publishing, 2014: 834-849.

[4] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry[C]//2014 IEEE International Conference on Robotics and Automation (ICRA), May 31-June 7, 2014, Hong Kong, China. New York: IEEE Press, 2014: 15-22.

[5] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 611-625.

[6] Saputra M R U, Markham A, Trigoni N. Visual SLAM and structure from motion in dynamic environments: a survey[J]. ACM Computing Surveys, 2018, 51(2): 1-36.

[7] Kundu A, Krishna K M, Sivaswamy J. Moving object detection by multi-view geometric techniques from a single camera mounted robot[C]//2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 10-15, 2009, St. Louis, MO, USA. New York: IEEE Press, 2009: 4306-4312.

[8] Migliore D, Rigamonti R, Marzorati D, et al. Use a single camera for simultaneous localization and mapping with mobile object tracking in dynamic environments[C]//ICRA Workshop on Safe Navigation in Open and Dynamic Environments: Application to Autonomous Vehicles. 2009: 12-17.

[9] Lin F C, Liu Y H, Zhou J F, et al. Optimization of visual odometry algorithm based on ORB feature[J]. Laser & Optoelectronics Progress, 2019, 56(21):

211507.
林付春, 刘宇红, 周进凡, 等. 基于 ORB 特征的视觉里程计算法优化[J]. 激光与光电子学进展, 2019, 56(21): 211507.

[10] Klappstein J, Vaudrey T, Rabe C, et al. Moving object segmentation using optical flow and depth information[C]// Advances in Image and Video Technology, 2009: 611-623.

[11] Lin Z L, Zhang G L, Yao E L, et al. Stereo visual odometry based on motion object detection in the dynamic scene[J]. Acta Optica Sinica, 2017, 37(11): 1115001.
林志林, 张国良, 姚二亮, 等. 动态场景下基于运动物体检测的立体视觉里程计[J]. 光学学报, 2017, 37(11): 1115001.

[12] Xiao L H, Wang J G, Qiu X S, et al. Dynamic-SLAM: semantic monocular visual localization and mapping based on deep learning in dynamic environment[J]. Robotics and Autonomous Systems, 2019, 117:1-16.

[13] Wang Z M, Zhang Q, Li J S, et al. A computationally efficient semantic SLAM solution for dynamic scenes[J]. Remote Sensing, 2019, 11(11): 1363.

[14] Brasch N, Bozic A, Lallemand J, et al. Semantic monocular SLAM for highly dynamic environments[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 1-5, 2018, Madrid, Spain. New York: IEEE Press, 2018: 393-400.

[15] Yu C, Liu Z, Liu X J, et al. DS-SLAM: a semantic visual SLAM towards dynamic environments[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.

[16] Zhao L L, Liu Z L, Chen J W, et al. A compatible framework for RGB-D SLAM in dynamic scenes[J]. IEEE Access, 2019, 7: 75604-75614.

[17] Canny J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(6): 679-698.

[18] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2020-07-10]. https://arxiv.org/abs/1409.1556.

[20] Gao X, Zhang T. 14 lectures on visual SLAM: from theory topractice[M]. Beijing: Publishing House of Electronics Industry, 2017.
高翔, 张涛. 视觉 SLAM 十四讲: 从理论到实践[M]. 北京: 电子工业出版社, 2017.

[21] Zhao L, Huang S D, Yan L, et al. Large-scale monocular SLAM by local bundle adjustment and map joining[C]//2010 11th International Conference on Control Automation Robotics & Vision, December 7-10, 2010, Singapore, Singapore. New York: IEEE Press, 2010: 431-436.

[22] Sturm J, Engelhard N, Endres F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]// 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 7-12, 2012, Vilamoura, Portugal. New York: IEEE Press, 2012: 573-580.