doi:10.6041/j.issn.1000-1298.2022.10.032

基于信息融合描述子的机器人复杂场景位姿估计算法

齐咏生[1,2]　姚辰武[1]　刘利强[1]　董朝轶[1,2]　李永亭[1,2]

(1.内蒙古工业大学电力学院,呼和浩特 010080;2.内蒙古自治区机电控制重点实验室,呼和浩特 010051)

摘要:传统机器人 V-SLAM 前端定位算法是基于人工设定的特征点提取和描述子局部匹配进行定位的,由于人工设定的主观性会导致提取方法鲁棒性差、复杂场景适应能力弱(场景明亮变化、噪声的引入、运动模糊)以及局部描述子匹配精度低等问题,为此,提出一种前端定位算法(SuperPoint Brief and K-means visual location, SBK-VL),该算法首先采用一种改进的概率 p-SuperPoint 深度学习算法提取特征点,以解决特征点鲁棒性低、复杂场景适应力弱的问题。其次提出一种全局信息(特征点聚类)和局部信息(Brief 描述子)相结合的复合描述子,降低传统描述子误匹配及匹配精度低的问题,实验结果显示该复合描述子的平均匹配正确率为 92.71%。最后将该 SBK-VL 替换 ORB-SLAM2 的前端,引入一种 Ransac 随机抽样方法对位姿进行检验,并使用绝对轨迹误差、相对轨迹误差、平均跟踪时间与 ORB-SLAM2 算法和 GCNv2-SLAM 算法进行比较。实验结果表明,本文算法具有更好的均衡性能,一方面可提升经典 V-SLAM 算法的复杂场景适应性和估计精度,另一方面相比传统深度学习 SLAM 算法具有更好的实时性和更低计算成本。

关键词:自主移动机器人;位姿估计;复杂场景;V-SLAM;SLAM 前端;深度学习

中图分类号:TP242.6　　**文献标识码**:A　　**文章编号**:1000-1298(2022)10-0293-13　　**OSID**:

Pose Estimation Algorithm for Robot Complex Scenes Based on Information Fusion Descriptor

QI Yongsheng[1,2]　YAO Chenwu[1]　LIU Liqiang[1]　DONG Chaoyi[1,2]　LI Yongting[1,2]

(1. *Institute of Electric Power, Inner Mongolia University of Technology, Huhhot* 010080, *China*
2. *Inner Mongolia Key Laboratory of Electrical and Mechanical Control, Huhhot* 010051, *China*)

Abstract: The traditional robot V-SLAM front-end positioning algorithm is based on manually set feature point extraction and descriptor local matching for positioning. Due to the subjectivity of manual setting, the extraction method will have poor robustness and weak adaptability to complex scenes (scene brightness changes, the introduction of noise, motion blur) and the low accuracy of local descriptor matching. For this reason, a front-end positioning algorithm (SuperPoint Brief and K-means visual location, SBK-VL) was proposed. The algorithm firstly used an improved p probability-SuperPoint deep learning framework extracted feature points to solve the problem of low robustness of feature points and weak adaptability to complex scenes; secondly, a combination of global information (feature point clustering) and local information (Brief descriptor) was proposed. Descriptors can reduce the mismatch of traditional descriptors and improve the problem of low matching accuracy. The experimental results showed that the average matching accuracy rate was 92.71%. Finally, replacing the SBK-VL with the front end of ORB-SLAM2, a Ransac random sampling method was used to test the pose, and the absolute trajectory error index was used. Relative trajectory error index and average tracking time were compared with that of ORB-SLAM2 algorithm and GCNv2-SLAM algorithm. The experimental results showed that the algorithm had better equalization performance. On the one hand, it can improve the complex scene adaptability and estimation accuracy of the classic V-SLAM algorithm. On the other hand, it had better real-time performance and computational cost than the traditional deep learning SLAM algorithm.

Key words: autonomous mobile robot; pose estimation; complex scene; V-SLAM; SLAM front end; deep learning

收稿日期:2021-11-24　修回日期:2022-04-19

基金项目:国家自然科学基金项目(61763037)、内蒙古自治区科技计划项目(2021GG164)和内蒙古自治区自然科学基金项目(2020MS05029、2021MS06018)

作者简介:齐咏生(1975—),男,教授,博士,主要从事机器人协同控制技术研究,E-mail:qys@imut.edu.cn

0 引言

为实现自主移动和导航,移动机器人需要解决3个关键问题:位姿估计、建图、路径规划[1-3]。其中移动机器人位姿估计是建图与路径规划的重要前提和关键环节。

目前常见的信息提取渠道主要分为激光雷达传感器和相机传感器。而相机传感器相比激光雷达传感器,具有成本低廉、质量轻、功耗低、采集信息丰富等优点,因而相机传感器用于机器人位姿估计已成为一个重要的发展方向。

即时定位与地图构建(Simultaneous localization and mapping,SLAM),主要解决移动机器人在未知环境中定位与地图构建的问题。而 SLAM 前端作为其重要环节是一个求解运动变换的过程,并为后端提供初始的数据,SLAM 前端可以分为特征提取、特征匹配、数据关联[4-6]。迄今为止已经有一些经典的视觉 SLAM 算法,其算法的前端部分都是采用视觉定位。MonoSLAM[7]是第一个基于单目相机运行的 SLAM 方案,其前端采取稀疏的特征点进行匹配估计位姿,能够实现实时的无漂移运动恢复结构。但该方案的前端特征点跟踪鲁棒性较差,很容易丢失;PTAM[8]采用了 Tracking 和 Mapping 双线程的结构,在 Tracking 线程中其前端采取 FAST 提取特征点进行匹配,其优点是提取速度快,但也存在鲁棒性差的缺陷,而且它会提取出大量的 FAST 特征,需要使用各种约束剔除误匹配,耗时耗力。为了降低前端计算工作量,以直接法为代表的 LSD-SLAM[9]在帧与帧之间实现跟踪以构建光度误差,降低特征匹配的计算量,但该算法对纹理缺乏的场景效果较差。ORB-SLAM2[10]在 PTAM 架构的基础上,增加了地图初始化和闭环监测功能,可实现全局约束优化。但 ORB-SALM2 前端在复杂场景下(噪声引入、运动模糊、光线明亮变化)特征点的提取抗干扰能力较差,而且其描述子 Brief[11]利用的是图像局部信息,缺乏全局性描述,会对提取的特征点造成较多误匹配,影响定位精度。

为进一步提高 ORB-SLAM2 定位精度,MA 等[12]从提高特征点描述子区分度的角度出发提出了 ASD-SLAM,针对视点变化和高度重复的场景,使用深度学习的方法提高特征点描述子的区分度。所提出的 Adaptive Scale Triplet 损失函数应用于 Triplet Network 获得自适应尺度描述子(ASD),替换传统视觉 SLAM 框架的前端,提高了特征的匹配精度。然而该算法也存在特征点提取抗干扰能力差,在复杂环境下提取点数变少甚至无法求解位姿的问题。LI 等[13]将深度学习目标检测方法融入 SLAM 前端中,提出一种 PO-SLAM 方案,该方法是对 ORB-SLAM2 的改进,对输入的图像从目标检测网络中提取对象的语义信息,使用 bundle adjustment 构建对目标的映射,可有效提高定位精度。不过,当环境中目标信息减少和缺失时,仍会出现目标检测失效、定位精度下降的问题。

为克服上述问题,提出基于多级深度网络提取图像特征点和输出描述子[14-17]。在面对复杂场景提取特征点时具有很好的抗干扰性,输出的描述子匹配精度也有较大提高。但上述网络对特征点提取和描述子输出分别采用不同网络,网络结构庞大复杂,当面对较大场景、输入图像较大时,网络的计算时间成本陡然增加,降低了 SLAM 系统的实时性,此外该类方法对移植系统的资源要求很高。

通过前期研究和深入分析,发现目前所研究的 ORB-SLAM2 的前端定位算法主要缺陷为:特征点的提取是基于人工设定的规则,在复杂场景[18-21]中,特征点提取会失效,或点数变少甚至于无法求解位姿,容易跟踪失败。通过使用深度学习可有效改善这些问题,然而引入深度学习在特征点提取和描述子输出时会带来网络庞大、复杂场景环境下计算成本过高,实时性难以保证的问题。ORB-SLAM2 使用 Brief 描述子利用的是图像的局部信息进行特征点匹配,而图像中纹理信息较为丰富,当特征点距离较远时局部信息描述子可能非常相似,很容易产生误匹配现象,影响定位精度。为此,从不影响 V-SLAM 实时性同时又保证具有较强抗干扰性角度出发,提出一种概率 p-SuperPoint 网络实现特征点提取,提高提取特征点的鲁棒性;并在其中引入一种带有反馈特征点 p 概率控制策略,实现特征点的输出质量控制,提高前端位姿估计的准确性。提出在描述子输出部分引入一种自适应 K-means 聚类算法融合 Brief 描述子,提高特征点的匹配精度和鲁棒性。最后,引入一种 Ransac 的随机抽样方法对估计的位姿进行检验,增强算法的可靠性。

综上,本文提出一种信息融合描述子位姿估计前端算法(SuperPoint Brief and K-means visual cocation,SBK-VL)。该算法综合考虑 ORB-SLAM2 和深度学习方法 2 种思路的优缺点,给出一种均衡方案,用更小的计算成本得到最佳的估计精度;将该算法与经典 ORB-SLAM2 算法、GCNv2-SLAM 算法进行比较,以验证算法的有效性。

1　理论基础

1.1　SuperPoint 特征点提取框架

SuperPoint 网络是基于自监督学习用于训练特征点检测器和描述子，适用于计算机视觉中大多数视点几何问题[14]。它能够有效检测图像中的特征点和输出特征点对应的描述子，其中 SuperPoint 网络中的特征点提取网络在复杂环境下具有较好的抗干扰能力。SuperPoint 特征点提取网络框架如图 1 所示。

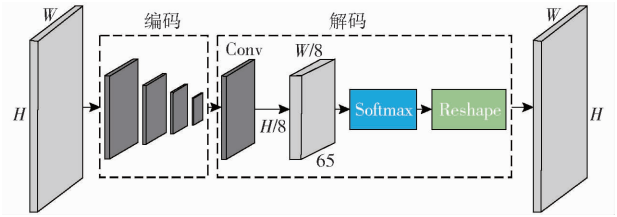


图 1　SuperPoint 特征点提取网络框架图

Fig. 1　SuperPoint feature point extraction network

该网络输入尺寸 $W \times H$ 图像，经过类似于 VGG－net 编码形式的卷积，对图像实施降维，提取特征，图中的解码部分经过 Softmax，计算图像的每个像素成为特征点的概率，最后在 Reshape 中采用子像素卷积的方法将其重组为与原来图像尺寸相同并带有特征点位置的图像。该网络输出图像中的像素带有一个 0－1 的数值概率表示该像素点成为特征点的概率[14]。该网络结构简单，参数组成较少，网络运算时间短，因此可以将其应用到 V－SLAM 算法中，来提取复杂环境下的特征点。

1.2　多视图几何技术

多视图几何中已知 2D 的像素坐标估计相机的运动使用的是对极几何。

如图 2 中 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是一对在第 1 帧和第 2 帧中匹配好的特征点，设第 1 帧到第 2 帧的运动为旋转和平移，旋转和平移运动前后的 2 个相机的中心分别为 O_1、O_2。设 $\boldsymbol{P}$ 的空间位置为 $\boldsymbol{P}=(X,Y,Z)$。

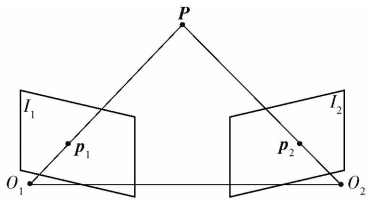


图 2　对极几何约束

Fig. 2　Polar geometric constraints

根据针孔相机原理，2 个像素点 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 的像素位置为

$$s_1\boldsymbol{p}_1=\boldsymbol{KP} \tag{1}$$

$$s_2\boldsymbol{p}_2=\boldsymbol{K}(\boldsymbol{RP}+\boldsymbol{t}) \tag{2}$$

式中　$\boldsymbol{K}$——相机内参矩阵

$\boldsymbol{R}$——相机旋转矩阵

$\boldsymbol{t}$——相机平移向量

s_1、s_2——尺度

上述 2 个投影关系可以写为

$$\boldsymbol{p}_1\cong\boldsymbol{KP} \tag{3}$$

$$\boldsymbol{p}_2\cong\boldsymbol{K}(\boldsymbol{RP}+\boldsymbol{t}) \tag{4}$$

式中≅表示尺度意义下的相等，现取

$$\boldsymbol{x}_1=\boldsymbol{K}^{-1}\boldsymbol{p}_1 \tag{5}$$

$$\boldsymbol{x}_2=\boldsymbol{K}^{-1}\boldsymbol{p}_2 \tag{6}$$

式中　$\boldsymbol{x}_1$、$\boldsymbol{x}_2$——2 个像素归一化平面的坐标

于是有

$$\boldsymbol{x}_2\cong\boldsymbol{R}\boldsymbol{x}_1+\boldsymbol{t} \tag{7}$$

对两边同时左乘 $\hat{\boldsymbol{t}}$，然后两侧同乘 $\boldsymbol{x}_2^{\mathrm{T}}$ 得

$$\boldsymbol{x}_2^{\mathrm{T}}\hat{\boldsymbol{t}}\boldsymbol{x}_2\cong\boldsymbol{x}_2^{\mathrm{T}}\hat{\boldsymbol{t}}\boldsymbol{R}\boldsymbol{x}_1 \tag{8}$$

式中　$\hat{\boldsymbol{t}}$——$\boldsymbol{t}$ 的反对称矩阵

化简后，得到用对极几何计算相机的位姿，计算式化简为

$$\boldsymbol{p}_2^{\mathrm{T}}\boldsymbol{K}^{-\mathrm{T}}\hat{\boldsymbol{t}}\boldsymbol{R}\boldsymbol{K}^{-1}\boldsymbol{p}_1=0 \tag{9}$$

式(9)的中间部分记作 2 个矩阵：基础矩阵 $\boldsymbol{F}$ 和本质矩阵 $\boldsymbol{E}$。

$$\boldsymbol{F}=\boldsymbol{K}^{-\mathrm{T}}\boldsymbol{E}\boldsymbol{K}^{-1} \tag{10}$$

$$\boldsymbol{E}=\hat{\boldsymbol{t}}\boldsymbol{R} \tag{11}$$

可以进一步化简对极约束，根据特征点的配对坐标求出 $\boldsymbol{E}$ 或 $\boldsymbol{F}$，根据 $\boldsymbol{E}$ 或 $\boldsymbol{F}$ 求出 $\boldsymbol{R}$、$\boldsymbol{t}$。

2　SBP－VL 算法框架

2.1　算法架构

SBP－VL 算法架构如图 3 所示。由图 3 可知，算法框架主要包括 3 部分：基于位姿反馈输出的深度学习特征点提取网络、特征点匹配和位姿估计。

(1)基于概率 p－SuperPoint 网络的特征点提取。如前所述，SuperPoint 网络编码器具有类似 VGG－net 的架构，其网络结构简单、参数少、运算时间短，非常适合应用到 V－SLAM 算法中实现复杂环境下特征点提取。但在应用中该网络需要提前设定一个固定的阈值来判定图中像素点是否为特征点。实验中发现，固定的阈值对于不同场景(如运动模糊场景)会导致连续丢帧，甚至是位姿估算失败进入重定模式等。为此，引入一种概率阈值 p 控制方法，即由特征点的概率阈值 p 自适应控制该网络生成特征点的质量，当位姿估算超出误差范围或位姿估算失败时特征点的概率阈值 p 会被重置，以适应场景的变化实现准确位姿估计。另外，由于网络输出的特征点可能会有稠密现象，给后续特征点匹配带来一定难度，因此，在本文算

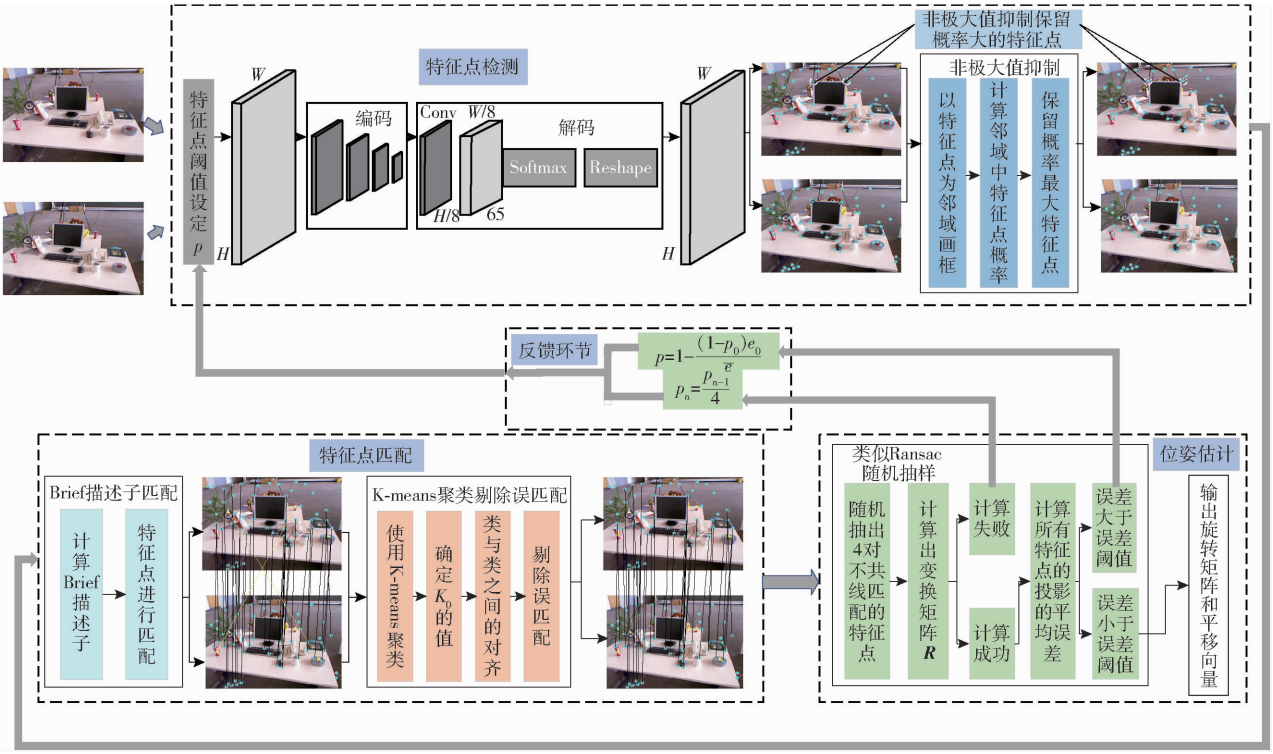


图 3 SBK－VL 算法框架

Fig. 3 SBK－VL algorithm framework

法中加入了非极大值抑制环节，将多余特征点去除。

(2)特征点匹配。该部分由 Brief 描述子匹配和特征点 K-means 聚类组成。考虑到传统 Brief 描述子主要反映的是局部特征，会造成误匹配。为此，基于相邻帧结构相似性原理，引入自适应 K-means 聚类剔除出现的误匹配，同时采用欧氏距离平均法自适应确定聚类类数 K_0。

(3)位姿估计。引入 Ransac 随机抽样对估计出的位姿进行准确性检验，设置误差检验范围并加入反馈环节。若估计的位姿结果超出该范围或位姿估计失败，则进行反馈重新设置概率阈值 p，重新产生特征点；否则，输出估计准确的旋转矩阵和平移向量。

2.2 特征点检测

SuperPoint 网络是一个能够检测特征点和描述子的双网络。但当输入图像变大、场景变大，该网络的计算成本增加、运算量变大。因此，在特征点检测中只使用 SuperPoint 网络中的特征点检测网络，这样可以减少该网络的计算时间。该特征点检测网络能够对图像中的像素点生成一个概率值的标记，表示该像素点成为特征点的概率，因此将 SuperPoint 网络输出的图像中像素点概率作为衡量特征点质量的标准。通过设定概率阈值 p 便可以控制特征点的输出。概率阈值 p 设置的越高表示网络提取到的特征点质量越高，输出数量越少。特征点输出质量越高表示特征点在不同场景下的不变性与稳定性越强。而前端的位姿估计面临的 2 个风险分别为由特征点误匹配造成的位姿估计误差太大和由特征提取数目太少而造成的位姿估计失败。而这 2 个问题可以通过控制特征点的输出来解决。

(1)位姿估计误差太大。该情况大多是由于特征匹配出现误差，而位姿估计时使用了错误的特征匹配数据，对于前端位姿估计误差太大，解决办法为提高特征点的输出质量，减少特征点的输出数量，即提高概率阈值 p，一方面阈值 p 的提高使得特征点数量减少，复杂环境下特征点的描述子相似性很高的点多数会消失，降低匹配难度，且在特征点匹配的过程中进一步减少了时间；另一方面特征点输出质量高、稳定性强，在不同场景下处于前一帧图像的特征点很大概率也会存在后一帧图像中，在特征点匹配的过程中更加容易精确寻找到匹配的特征点，使得点与点匹配后的精度提高。对于前端位姿估计误差太大情况，本文提出概率阈值 p 的控制策略公式

$$p=\begin{cases}p_0 & (\bar{e}\leqslant e_0)\\ 1-\dfrac{(1-p_0)e_0}{\bar{e}} & (\bar{e}>e_0)\end{cases} \tag{12}$$

式中 p_0——初始设定阈值

$\bar{e}$——所有特征点映射平均误差

e_0——所有特征点映射误差阈值

误差阈值是以真实投影特征点为圆心，半径为

e_0 的圆，该误差阈值与图像的尺寸有一定的比例关系，即

$$\frac{\pi e_0^2}{WH}=0.002 \tag{13}$$

式中　W——输入图像的宽

　　H——输入图像的高

式(13)表示以真实投影特征点为圆心，半径为 e_0 的圆与输入图像的面积占比为 0.002，根据实际输入图像的尺寸可以确定 e_0。在实际场景中可以根据实际情况对 e_0 进行微调。

图 4 表示实际投影的 2 种情况：I_1 图像中的点 u_1、u_2，经过三维空间中点 P_1、P_2 映射到图像 I_2 中点 u_1'、u_2'，图像 I_2 中的虚线方框表示实际投影位置，实线方框为真实投影点，圆圈表示投影误差范围，该图展示了实际投影在误差范围内和超出误差范围的 2 种情况。

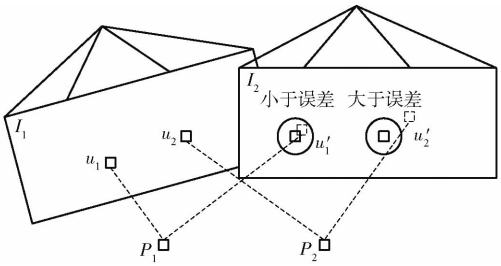


图 4　特征点投影示意图

Fig. 4　Projection of feature points

当位姿估计检验判定所有特征点映射的平均误差 $\bar{e}$ 小于误差阈值 e_0 时，概率阈值保持初设定 p_0 不变，当误差阈值大于 e_0 时根据式(12)计算出新的概率 p 重新提取特征点，该控制策略示意图和流程如图 5、6 所示。

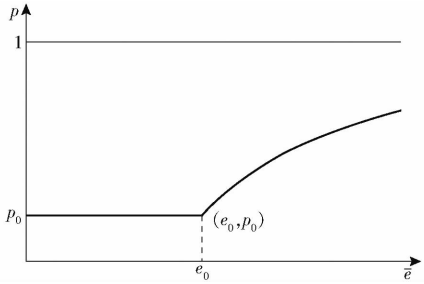


图 5　控制策略示意图

Fig. 5　Schematic of control strategy

(2)位姿估计失败。当复杂场景中提取到的特征点数目太少(如运动模糊)而无法满足位姿方程的求解条件，往往会造成位姿估计失败，针对这种情况，解决办法为降低概率阈值 p、提高特征点数量、降低质量以满足位姿求解方程。对于位姿估计失败时该概率阈值 p 的控制策略公式为

$$p_n=\frac{p_{n-1}}{4} \tag{14}$$

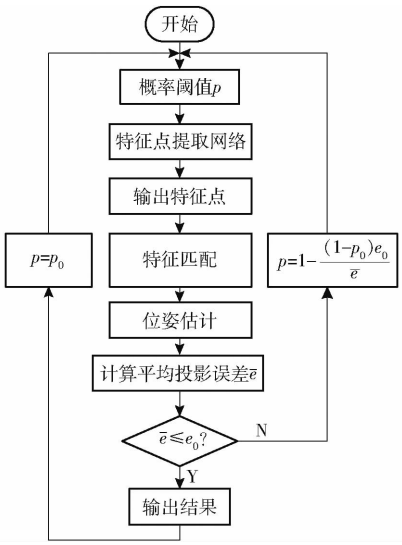


图 6　阈值重设流程图

Fig. 6　Threshold reset flow chart

式中　p_{n-1}——$n-1$ 次设置的概率阈值

　　p_n——n 次设置的概率阈值

该控制策略采用比例函数，当 $n-1$ 次设置的概率阈值 p_{n-1} 使得提取的特征点数量不足造成位姿求解失败时，为保证降低后概率阈值所提取出的特征点数能够满足式(11)的求解(式(11)的求解需要至少 4 对匹配的特征点)，所以降低后的概率阈值所提取出的特征点数应至少为原来的 4 倍，于是概率阈值设置降低为原来的 1/4，重新提取特征点求解位姿。该控制策略流程如图 7 所示。

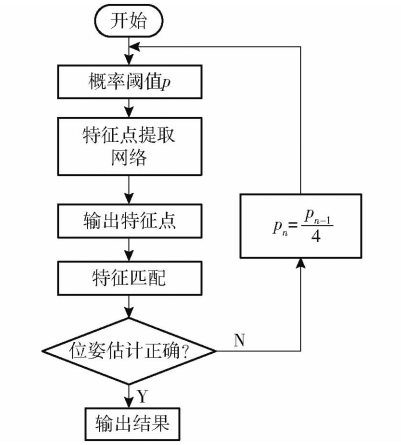


图 7　位姿估计失败时阈值重设流程图

Fig. 7　Threshold reset flow chart when pose estimation failed

2.3　特征点匹配

特征点匹配如图 8 所示，包括特征点预匹配、匹配特征点的自适应 K-means 聚类、类与类之间的对齐剔除误匹配 3 部分。

2.3.1　特征点预匹配

将前后 2 帧检测的特征点使用 Brief 描述子进行预匹配。由于 Brief 描述子是一种利用图像局部信息产生的描述子，且采用二进制描述，因此会造

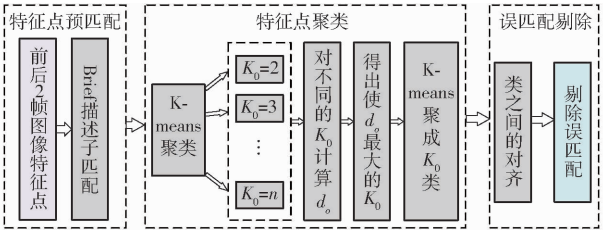


图 8 特征点匹配算法流程图

Fig. 8 Feature point matching algorithm flow

成特征点误匹配现象。此时即使降低阈值也不能消除误匹配,这是因为图像中的纹理信息丰富,使得特征点的局部信息很容易产生相似性,而阈值的下降会造成点数减小,导致方程求解条件缺失无法求解。因此必须进行进一步误匹配消除。

2.3.2 特征点聚类

在特征点进行预匹配的过程中,发现经过预匹配的特征点在结构上具有相似性,因此当对参与预匹配的特征点进行聚类时,前后 2 帧图像的特征点能够有很好的聚类性能,可充分利用前后 2 帧类的相似性剔除误匹配。基于此,提出 K-means 聚类的误匹配剔除思路。使用 K-means 聚类主要是因为该方法计算简单、直观,不过传统 K-means 聚类算法需要人为设定聚类系数 K_0,实际应用中很难满足,为此,提出一种自适应的改进措施。

首先定义距离

$$d_0 = \frac{\sum_{i=1}^{K_0}\sum_{j=i+1}^{K_0}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\frac{K_0(K_0 - 1)}{2}} \tag{15}$$

式中 i、j——图像中特征点聚类数

(x,y)——聚类的中心点坐标

d_0——类间的平均欧氏距离

当该距离越大时,其一帧图像中类与类之间距离越远,在进行类与类的对齐时越易于匹配。

选取不同的 K_0 则会对应不同 d_0,d_0越大,其类与类之间的差距越大则特征点的结构信息越明确,2 帧图像类与类之间的匹配范围就越清晰,因此 K_0 的确定就转换为最大化距离 d_0问题。通过实验发现大部分的 K_0 取 3 和 4 时 d_0能够取得最大,但 K_0 的选取范围不能太大,过大的 K_0 会使得 K-means 在聚类时花费更多的时间,因此,将 K_0 的合理范围划分在 2 ~ 9 之间,能够满足特征点类的划分和适当的计算成本。由此得到具体算法流程如图 9 所示。

2.3.3 误匹配剔除

误匹配剔除的关键在于 2 帧图像中第 1 帧图像的类与第 2 帧图像类的对齐,而类的中心点可以表

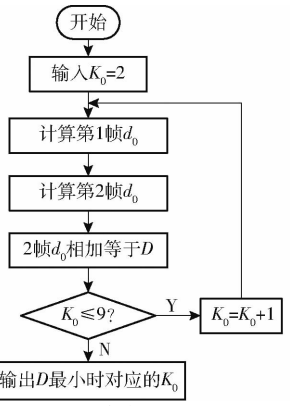


图 9 确定 K_0 流程图

Fig. 9 Flow chart for determining K_0 value

示一类特征点的范围,因此类与类的对齐就转换为中心点匹配问题。首先,对输入的前后 2 帧图像根据图 9 流程图所确定的 K_0 分别聚成 K_0 类;其次计算前后 2 帧 K_0 类的中心点坐标;在类的中心点周围以 9 ×9 的方格取 81 位像素值作为该中心点的描述子,具体示意图如图 10 所示。

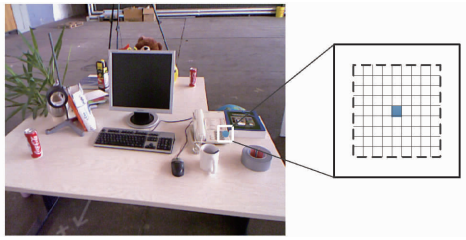


图 10 类中心点的描述方式

Fig. 10 Description of class center point

图 10 中的实心点表示类中心点,网格表示类中心点周围像素值。对每个中心点分别取上述描述方式,并在 2 帧之间进行匹配,与预匹配结果进行比较,最终剔除误匹配。显然类的中心点的对齐代表的是 2 类特征点结构的相似性和全局匹配性,可从整体范围内剔除特征点的误匹配,充分利用了全局信息。为了更好地说明误匹配剔除过程,示例如图 11 所示。

图 11 中实心圆点表示 Brief 描述子匹配后的特征点;三角形的点和虚直线表示使用 Brief 所产生的误匹配;正方形点表示聚类的中心点,虚线圆圈表示特征点聚类Ⅰ的范围,实线圆圈表示特征点聚类Ⅱ的范围。由图 11 可以看出,本该在类Ⅰ的特征点被 Brief 描述子匹配到了类Ⅱ的范围里,发生误匹配,为此采用聚类中心匹配方法,利用前后 2 帧图像类Ⅰ和类Ⅱ的划分和对齐可以有效地将误匹配结果剔除。

2.4 位姿估计

随机抽样一致性 Ransac,能够采用迭代的方式从一组被观测数据中估算出数学模型的参数。在经过

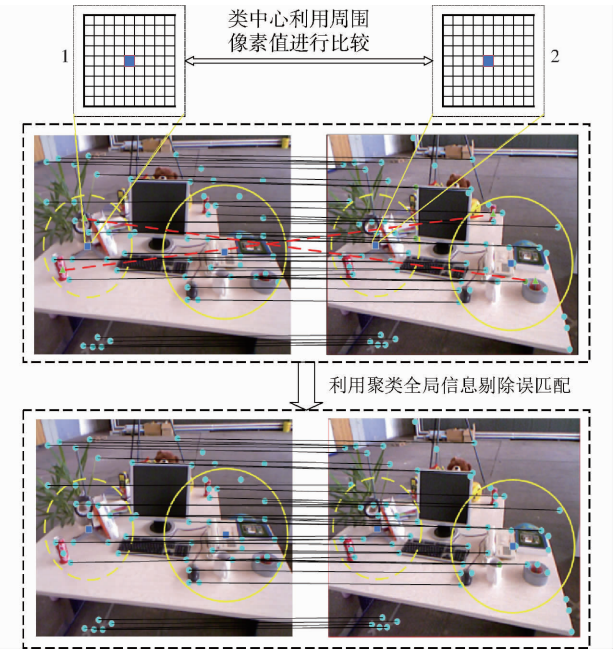


图 11　类的匹配示意图

Fig. 11　Schematic of matching

特征点匹配后利用多视图几何估计位姿，为了保证位姿估计的可靠性，使用一种基于 Ransac 随机采样方法的位姿准确性判定准则，该方法如下：首先从匹配好的特征点随机抽取 4 对特征点，将其称为内点，若该内点不足以构成求解方程条件则将原来的概率阈值 p 降低为原来的 1/4；反之利用内点计算出旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$；其次根据该变换矩阵计算除内点外所有特征点的投影误差并取平均值为 $\bar{e}$；若该误差平均值大于设定误差阈值 e_0 则通过式(12)重设特征点概率阈值 p；反之，则输出旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$。该位姿检验流程图如图 12 所示。

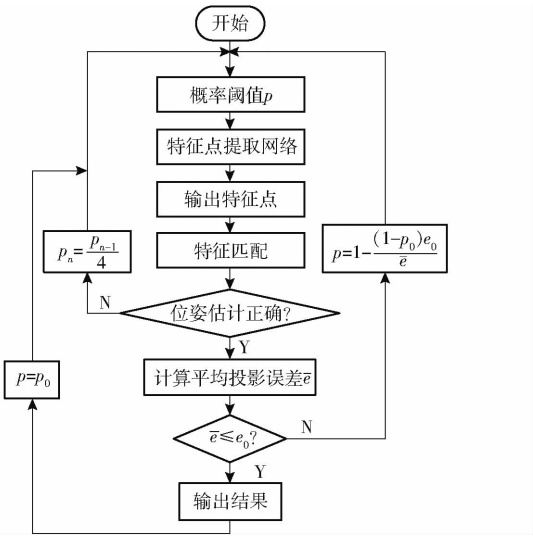


图 12　位姿检验流程图

Fig. 12　Flow chart of pose estimation

3　实验

为了验证本文算法的有效性，分别比较了 SuperPoint 与 FAST 特征点的场景适应性、概率阈值 p 控制策略正确性、提高概率阈值 p 对特征点匹配准确率的影响、Brief 描述子与 K-means 结合对特征点误匹配剔除的有效性等，最后，将本文方法与经典 ORB－SLAM2 算法[10]、GCNv2－SLAM 算法[16]进行了比较，验证算法的有效性和优越性。

3.1　SuperPoint 与 Fast 特征点的场景适应性

分别在明亮变化场景、运动模糊场景、引入高斯噪声和椒盐噪声场景，通过 SuperPoint 网络和 FAST 提取特征点，对比 2 种特征提取器的效果，实验结果如图 13 所示。

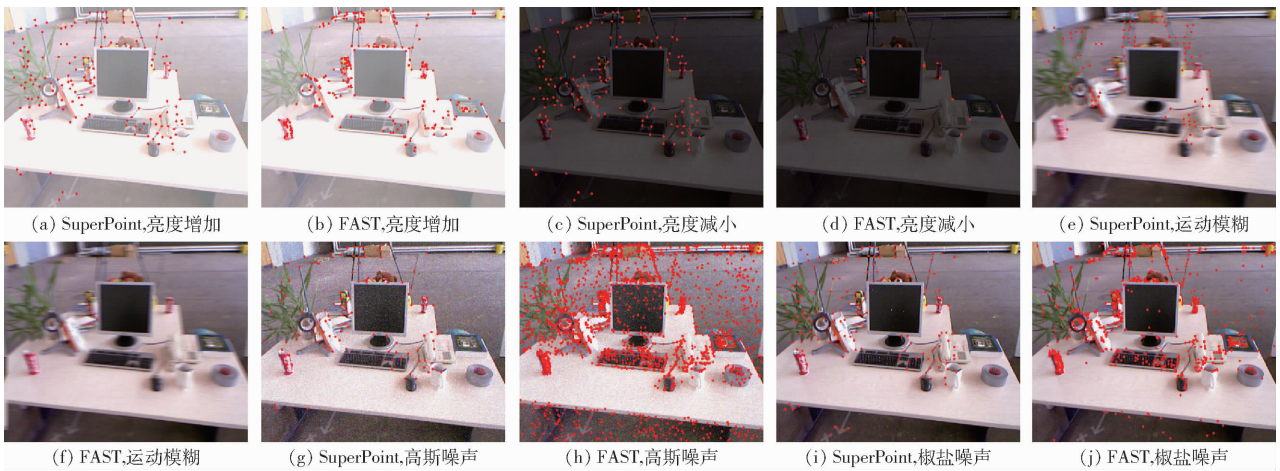


图 13　不同场景中两种算法对比结果

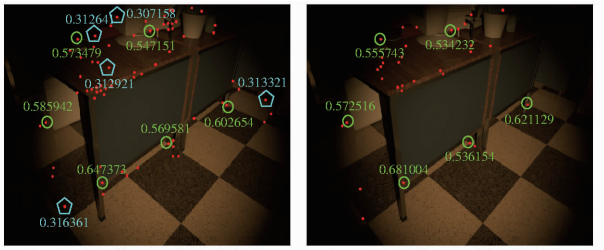
Fig. 13　Comparison results of two algorithms in different scenarios

图像中实心圆点表示提取到的特征点，由上述实验结果可以看出：在亮度减小时（夜间场景）SuperPoint 提取结果明显优于 FAST，FAST 提取点数不足以求解方程；在亮度增加时 SuperPoint 提取的特征点分布性较 FAST 好。在运动模糊场景下，SuperPoint 提取结果明显优于 FAST，FAST 提取点数

仅有 4 点。在对图像引入噪声后,发现 FAST 对噪声很敏感,而 SuperPoint 对噪声有抗干扰作用。综上,SuperPoint 的提取效果明显优于 FAST,对场景的适应能力更强。

3.2 概率阈值 p 的控制策略实验

SuperPoint 网络输出图像中的像素带有一个 0-1 的概率阈值标记,表示该像素点成为特征点的概率[14]。该概率阈值越高表示该特征点响应越高。如果设置一个概率阈值 p,使得网络输出的特征点大于概率阈值 p 的特征点,输出便筛选出了具有高概率的特征点。对前后 2 帧采用不同的概率阈值 p 进行特征点的提取,实验结果发现,带有高概率标记的特征点能够同时出现在前后 2 帧图像中,在帧与帧的变化过程中表现相对稳定,比带有低概率特征点稳定,不容易丢失,即带有高概率标记的特征点在应对环境变化时表现稳定,特征点质量高。图 14 为前后 2 帧带有概率标记的特征点图像。



(a) 概率阈值p=0.3提取特征点 (b) 概率阈值p=0.4提取特征点

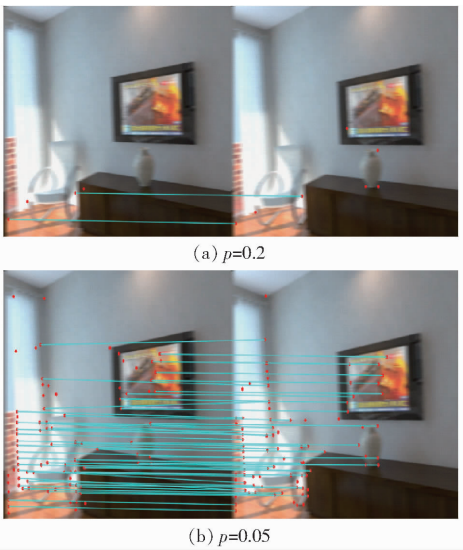
图 14 前后 2 帧带有概率标记的特征点图像

Fig. 14 Feature point images with probability markers in two frames before and after

图像第 1 帧采用概率阈值 $p=0.3$ 进行特征点的提取,第 2 帧图像采用概率阈值 $p=0.4$ 进行提取。其中使用圆圈标记的特征点显示有较高的概率标记值,也都存留在了第 2 帧中,而带有五边形标记的特征点显示有相对较低的概率标记值,在后一帧中没有保留。从上述结果可以发现,带有高概率标记的特征点在前后 2 帧中表现稳定。因此,采用 SuperPoint 网络提取特征点的输出响应值(概率阈值)P_i作为衡量该特征点质量,输出响应值 P_i越大表明质量越高,输出响应值 P_i越小表明质量越低。设定 p 为概率阈值,则集合 A 表示特征点质量相对高的集合,为 $A=\{P_i \mid P_i>p, 0<p<1\}$。

概率阈值 p 的改变有两种情况,分别为位姿估计失败和位姿估计误差太大,针对 2 种不同情况分别调整概率阈值 p 控制特征点的数量,设计 2 种实验。在 2 种实验中初始阈值 p_0 均设定为 0.2,实验采取的图像宽度为 480 像素,高度为 640 像素,根据式(13)可得误差阈值 e_0 约为 13.98 像素,为方便计算微调设定为 $10\sqrt{2}$像素。

对于位姿估计失败情况,在复杂环境中往往由于运动模糊产生的特征点输出较少,不满足位姿求解方程,针对这种情况,应当降低概率阈值产生更多的特征点,因此根据式(14)将初始阈值重置为原来的 1/4,即 0.05 重新提取特征点进行位姿估计,由图 15 可以看出概率阈值改变后特征点数明显增加,可进行位姿估计。



(a) p=0.2

(b) p=0.05

图 15 模糊场景下特征点提取匹配结果

Fig. 15 Feature point extraction and matching results when in fuzzy scene

对于位姿估计误差太大情况,其发生原因主要是由于特征匹配产生误差,而求解位姿方程时使用了该误差数据。对此,提高概率阈值减少特征点数量,数量减少,特征点的描述子相似性很高的点多数会消失以降低匹配难度;同时需要匹配的数目减少,节约了特征点匹配时间。依据策略,首先计算所有特征点映射的平均误差 $\bar{e}$,当平均误差大于误差阈值($10\sqrt{2}$像素)时根据式(12)将初始概率阈值 p_0 重置。以下选取了视频中的前后 2 帧来进行实验验证,在初始阈值 p_0 为 0.2 时,对前后 2 帧进行特征提取、特征匹配、位姿估计。计算出位姿估计误差为 17.86 像素,大于 $10\sqrt{2}$像素,因此依据式(12)计算得出阈值为 0.366 5,将初始阈值保留两位小数重置为 0.37。概率阈值重置后重新进行特征点提取、匹配、位姿估计。实验结果如图 15、16 所示。

图 16a 为 $p=0.2$ 时特征点的特征提取匹配情况,根据其特征匹配计算出 $\boldsymbol{R}$,当图 I_1 用 $\boldsymbol{R}$ 进行映射时并不能恢复图 I_2。对于图 16a 由于产生错误的位姿估计导致图像的映射结果错误,在图 16b中将 p 重置为 0.37,减少特征点数量提高特征点质量,计算出 $\boldsymbol{R}$ 能够很好地使其恢复至原图 I_2。

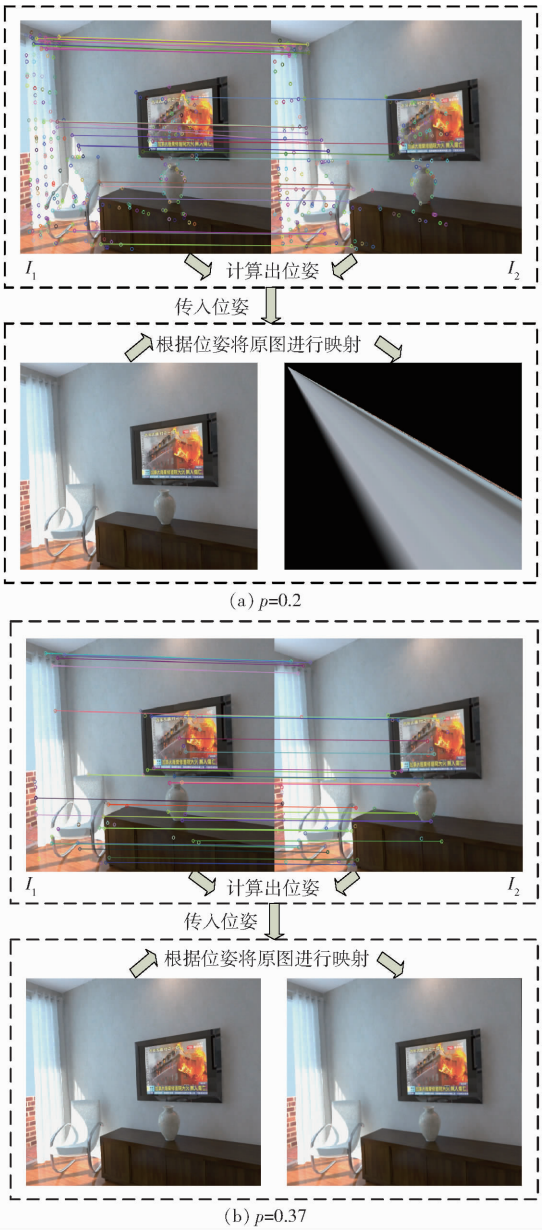


图 16　特征点提取匹配映射结果

Fig. 16　Feature point extraction matching mapping results

3.3　提高概率阈值对特征点匹配准确率的影响

为了验证提高特征点质量(即提高概率阈值)对特征点匹配准确率的影响,进行如表 1 所示实验,实验以概率阈值从 0.2 提升到 0.25 时特征点的匹配情况为例,并分别针对 TUM 和 ETHl 数据集进行测试,具体实验结果见表 1,图 17 为其中 1 组对比图结果。

图 17 对应实验 5(表 1 第 5 行),采用 ETHl 数据集进行验证结果。从对比图中可清晰地看出,在提高概率阈值后,原本在前一帧中存在的特征点(图中黄色圆圈标记)在下一帧中发生部分消失,此时由表 1 统计数据可知,消失的特征点中,误匹配点消失的比率更高。如实验 5 中,删除特征点总数为 30 个,其中误匹配特征点为 11 个,占比 36.7%,明显高于概率阈值提升之前的 19.5%。分析原因可知,误匹配的特征点质量偏低,当适当提升概率阈值后首先被剔除的特征点是这些低质量的特征点,因此本文算法相当于起到一个滤除作用,可以有效提升匹配的正确率。从表 1 也可以看出,对于给出的 6 个实验中,均实现了正确率的提升,平均提升率为 4.34 个百分点。由此,可以进一步表明,提高概率阈值有利于提升特征点的匹配正确率。

3.4　描述子与 K-means 结合对特征点误匹配剔除

首先使用特征点提取器对前后 2 帧图像进行特征点提取,之后使用 Brief 描述子进行匹配,在前后 2 帧中分别求 d_0,求出使 2 帧 d_0之和最大的值所对应的 K_0,确定聚类类数 K_0。由图 18 可以看出,当 $K_0=3$ 时取得最大值。将前后 2 帧图像分别分成 3 类,对分成 3 类的中心点进行对齐,剔除产生的误匹配。

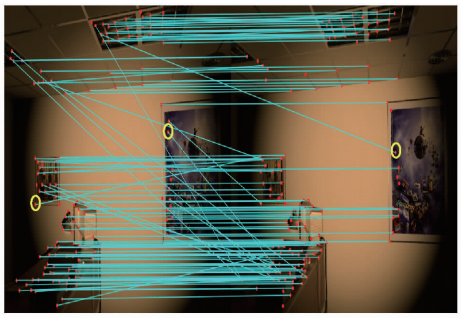
实验结果如图 19 所示,在 Brief 预匹配后特征点之间产生了误匹配,对参与匹配的特征点经过 $K_0=3$ 的聚类,因参与匹配的特征点在结构上具有相似性,利用其结构分布相似性对聚类后的特征点类进行中心点对齐,将帧与帧之间的类进行对齐。实验结果最后显示误匹配的特征点被剔除。

为了验证该匹配方法的准确程度,分别在 TUM 数据集和 ETHI 数据集中选取了几对图像帧进行了特征匹配实验,实验结果表明信息融合描述子的平

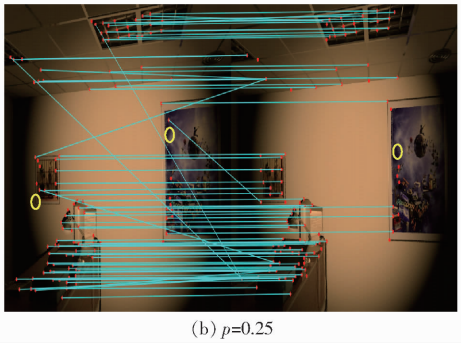
表 1　概率阈值提高后匹配情况

Tab. 1　Improved probability threshold matching

数据集	帧数	概率阈值 0.2 参与 Brief 匹配点对	概率阈值 0.2 正确匹配点对	概率阈值 0.2 正确率/%	概率阈值 0.25 时参与 Brief 匹配点对	概率阈值 0.25 正确匹配点对	概率阈值 0.25 正确率/%	正确率变化量/百分点
Tum-rgbd_freiburg1_desk	13&14	83	71	85.54	65	58	89.23	3.69
Tum-rgbd_freiburg1_xyz	120&121	104	84	80.77	89	76	85.39	4.62
Tum-rgbd_freiburg2_desk	150&151	117	101	86.32	108	99	91.66	5.34
Ethl-ethl_syn1	704&705	105	89	84.76	86	76	88.37	3.61
Ethl-ethl_syn1_flash	857&869	113	91	80.53	83	72	86.74	6.21
Ethl-ethl1_global	95&96	109	91	83.48	86	74	86.04	2.56



(a) p=0.2



(b) p=0.25

图 17 Brief 的匹配结果(Eth1-ethl_syn1_flash)

Fig. 17 Brief matching results

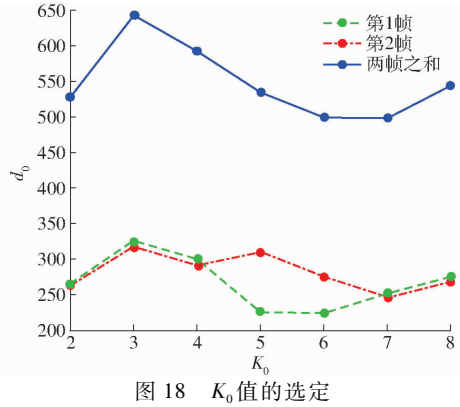


图 18 K_0 值的选定

Fig. 18 Selection of K_0 value

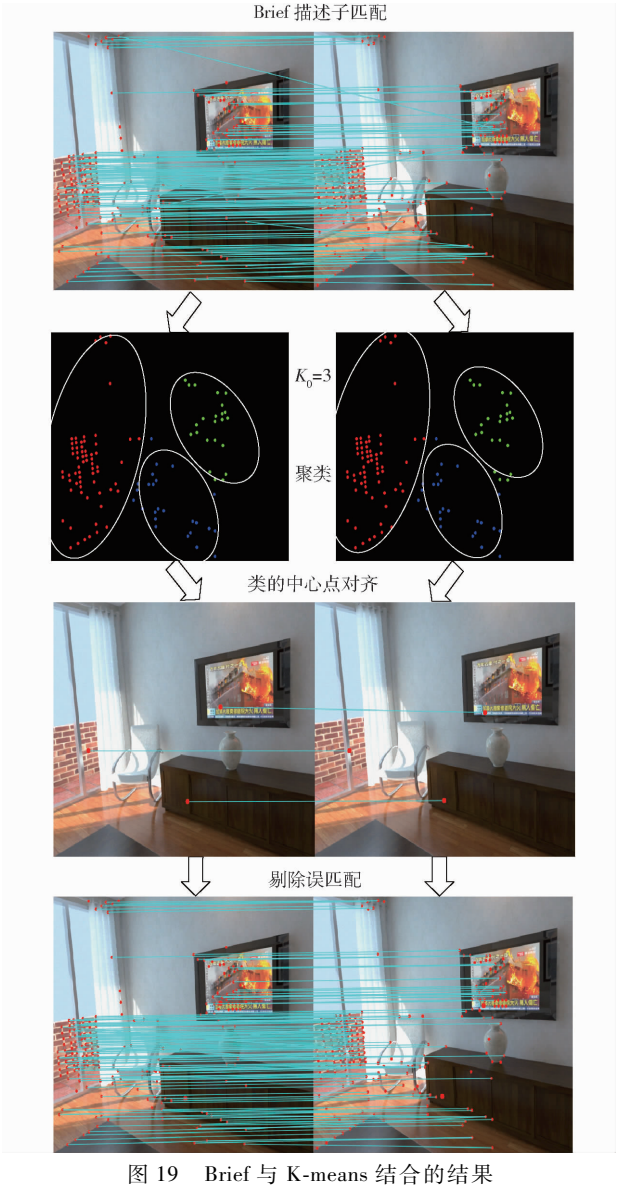


图 19 Brief 与 K-means 结合的结果

Fig. 19 Result of combination of Brief and K-means

均匹配正确率为 92.71%,Brief 描述子进行匹配的平均正确率为 80.86%。使用信息融合描述子对比使用 Brief 描述子的匹配正确率平均提高 11.87 个百分点。表 2 为该方法的匹配结果。

表 2 特征点匹配精度比较结果

Tab. 2 Feature point matching accuracy comparison

数据集	帧数	Brief 匹配点对	Brief 正确匹配点对	Brief 正确率/%	信息融合描述子匹配点对	信息融合描述子正确匹配点对	信息融合描述子正确率/%	正确率变化量/百分点
Tum-rgbd_freiburg1_desk	82&83	80	69	86.25	72	69	95.83	9.58
Tum-rgbd_freiburg2_desk	12&13	99	84	84.84	79	77	97.46	12.62
Tum-rgbd_freiburg3_walking_xyz	186&187	70	59	84.28	52	51	98.07	13.79
Tum-rgbd_freiburg1_xyz	1&2	110	93	84.54	98	91	92.85	8.31
Ethl-ethl1	857&869	99	73	73.73	77	66	85.71	11.98
Ethl-ethl_syn1	558&559	91	80	87.91	79	73	92.41	4.50
Ethl-ethl1_global	558&559	98	65	66.32	69	58	84.06	17.74
Ethl-ethl_syn1	558&570	88	68	77.27	52	49	94.23	16.96
Ethl-ethl_syn2-ethl2	907&908	46	38	82.61	32	30	93.75	11.14

3.5　位姿精度性能评价

为验证本文前端算法替换 ORB－SLAM2 前端能够提高定位精度，分别使用 TUM RGB－D 和 ETHl[22]公共数据集的多个典型视频序列对该算法进行测试，与 ORB－SLAM2[10]、GCNv2－SLAM 算法[16]在绝对轨迹误差（ATE）、相对轨迹误差（RPE）、平均跟踪时间等指标进行了比较。实验平台的硬件环境 CPU 为 AMD Ryzen 5 3600 CPU 3.60 GHz，GPU 为 NVIDIA GTX 1650 SUPER 4 GB，运行内存为16 GB。实验结果表明该算法能以每个序列的帧速率实时运行。基于 TUM 和 ETHl 的视频序列的实验结果如图 20～23 所示。

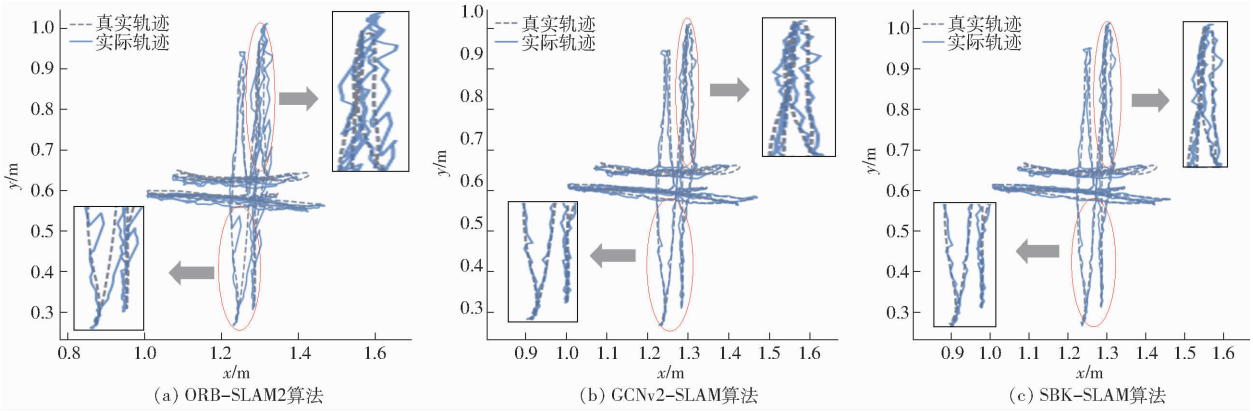


图 20　视频序列 fr1_xyz 不同算法仿真结果

Fig. 20　Simulation diagrams of different algorithms of video sequence fr1_xyz

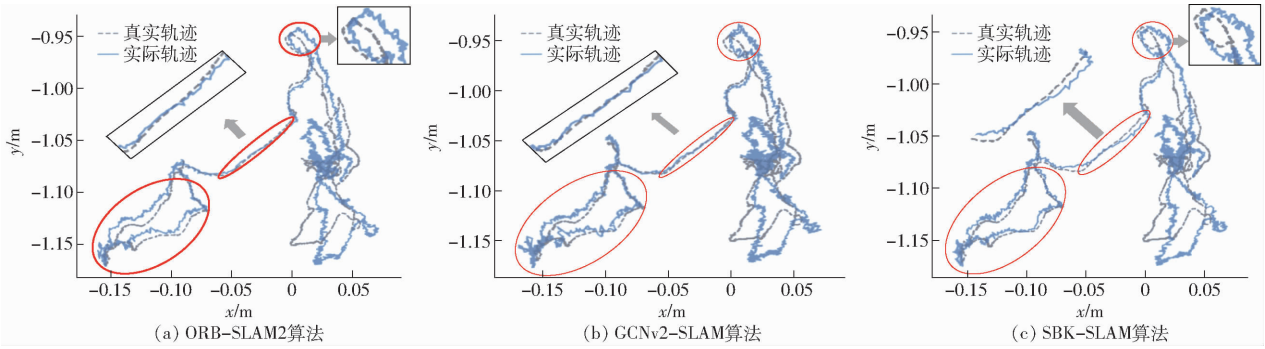


图 21　视频序列 fr2_rpy 不同算法仿真结果

Fig. 21　Simulation diagrams of different algorithms of video sequence fr2_rpy

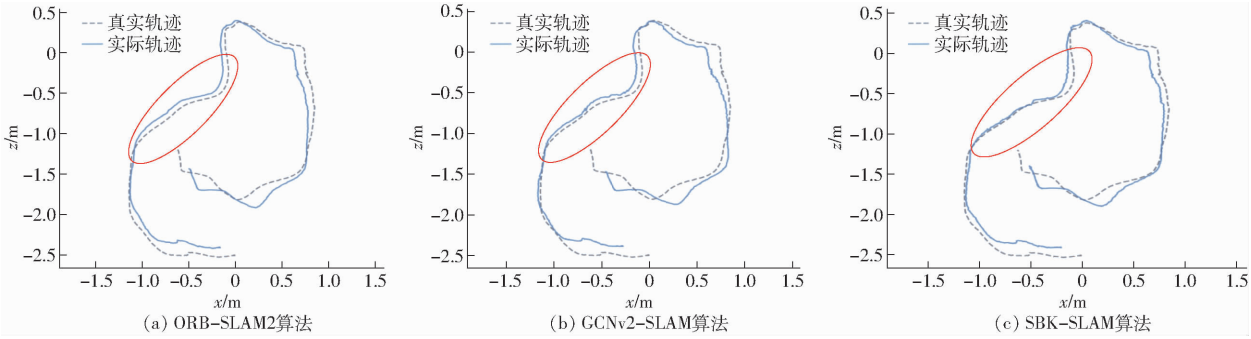


图 22　视频序列 Ethl_flash 不同算法仿真结果

Fig. 22　Simulation diagrams of different algorithms of video sequence Ethl_flash

图 20～23 实验结果展示了 3 种算法在 TUM 视频序列 fr1_xyz、fr2_rpy 和 ETHl 视频序列 Ethl_flash、Ethi_local 的测试结果，表 3、4 显示了 ETHl 和 TUM 视频序列在 3 种算法下的测试结果。

其中在 TUM 数据集下，在绝对误差和相对误差方面总体上 GCNv2－SLAM 最小，SBK－SLAM 次之，ORB－SLAM2 最大；而在 ETHl 数据集下，总体上 SBK－SLAM 最小，GCNv2－SLAM 与 ORB－SLAM2 相当。分析可知，当使用 TUM 数据集时 GCNv2－SLAM 的深度网络作用精度强使得误差最小，精度总体上均高于 SBK－SLAM、ORB－SLAM2。而当使用有干扰的数据集 ETHl 进行实验时，SBK－SLAM 精度总体上均高于 GCNv2－SLAM 和 ORB－SLAM2，分析可知主要在于 SBK－SLAM 在特征点

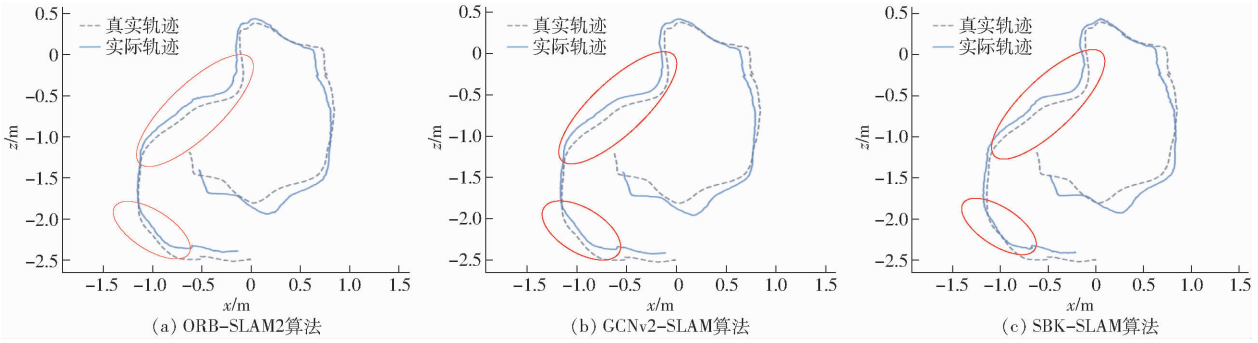


图 23 视频序列 Ethl_local 不同算法仿真结果

Fig. 23 Simulation diagrams of different algorithms of video sequence Ethl_local

表 3 ETHl 数据集下 SLAM 算法比较结果

Tab. 3 Comparison results of SLAM algorithms under ETHl dataset

视频序列	ORB－SLAM2 算法			GCNv2－SLAM 算法			SBK－SLAM 算法		
	ATE/m	RPE/m	平均跟踪时间/s	ATE/m	RPE/m	平均跟踪时间/s	ATE/m	RPE/m	平均跟踪时间/s
Ethl_syn1	2.441	0.020	0.016 8	2.461	0.016	0.026 8	2.419	0.014	0.019 8
Ethl_flash	2.459	0.018	0.016 8	2.452	0.013	0.023 8	2.421	0.017	0.020 3
Ethl_global	2.429	0.019	0.016 9	2.478	0.014	0.026 3	2.436	0.023	0.020 6
Ethl_loc_glo	2.443	0.019	0.017 3	2.426	0.013	0.026 8	2.429	0.012	0.019 9
Ethl_local	2.446	0.019	0.016 5	2.447	0.014	0.026 9	2.423	0.013	0.021 1

表 4 TUM 数据集下 SLAM 算法比较结果

Tab. 4 Comparison results of SLAM algorithms under TUM dataset

视频序列	ORB－SLAM2 算法			GCNv2－SLAM 算法			SBK－SLAM 算法		
	ATE/m	RPE/m	平均跟踪时间/s	ATE/m	RPE/m	平均跟踪时间/s	ATE/m	RPE/m	平均跟踪时间/s
fr1_desk	2.581	0.014	0.019 8	2.659	0.013	0.031 1	2.463	0.009	0.019 3
fr1_xyz	2.019	0.013	0.018 7	1.921	0.013	0.031 2	2.005	0.012	0.020 1
fr2_desk	2.981	0.008	0.018 9	2.560	0.005	0.045 0	2.718	0.007	0.032 0
fr2_rpy	1.886	0.008	0.015 8	1.882	0.002	0.036 1	1.886	0.005	0.027 0
fr2_xyz	2.067	0.009	0.016 4	1.995	0.005	0.036 3	2.026	0.008	0.019 3
fr3_str_tex_far	2.403	0.015	0.018 3	2.357	0.009	0.030 1	2.362	0.014	0.022 1

匹配期间除了计算 Brief 描述子外加入了 K-means 聚类算法,从而降低了误匹配使得最后的定位精度提高。

但在平均跟踪时间总体上 GCNv2－SLAM 用时最长、SBK－SLAM 次之、ORB－SLAM2 最小。主要原因在于 GCNv2－SLAM 特征点提取阶段和特征点匹配阶段使用了深度网络,SBK－SLAM 只在特征点提取阶段使用了深度网络,匹配阶段没有引入深度网络计算。

从总体看,SBK－SLAM 在匹配中增加了聚类算法,相比 ORB－SLAM2 提高了定位精度,在深度学习网络中只提取特征点适应场景复杂性,相比 GCNv2－SLAM 减少了运算时间。

4 结论

(1)在特征点提取阶段给出一种概率 p－SuperPoint 深度学习算法,相比于 ORB－SLAM2 的 FAST 特征提取器能更好地适应场景的复杂性,具有更高的鲁棒性。

(2)在特征点匹配阶段提出一种基于 Brief 描述子与自适应 K-means 聚类复合算法,实现局部细节和全局描述信息融合,可有效减少特征点的误匹配。

(3)在位姿估计阶段,使用一种 Ransac 随机抽样对估计位姿进行准确性检验,通过设置误差检验范围并加入反馈环节,提升位姿估计的准确性,增强算法的可靠性。

参 考 文 献

[1] LI A, RUAN X, HUANG J, et al. Review of vision-based simultaneous localization and mapping[C]// 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2019.

[2] 周俊,何永强.农业机械导航路径规划研究进展[J].农业机械学报,2021,52(9):1 - 14.
ZHOU Jun,HE Yongqiang. Research progress on navigation path planning of agricultural machinery[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(9):1 - 14. (in Chinese)

[3] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age[J]. IEEE Transactions on Robotics, 2016, 32(6):1309 - 1332.

[4] 齐咏生,陈培亮,刘利强,等.基于单目视觉的动态环境同步定位与多地图构建算法[J].农业机械学报,2022,53(4):280 - 292.
QI Yongsheng, CHEN Peiliang, LIU Liqiang, et al. Simultaneous localization and multi-mapping algorithm in dynamic environment based on monocular vision[J]. Transactions of the Chinese Society for Agricultural Machinery,2022,53(4):280 - 292. (in Chinese)

[5] HUANG Shoudong. A review of optimisation strategies used in simultaneous localisation and mapping[J]. Journal of Control and Decision, 2019, 6(1): 61 - 74.

[6] FUENTES-PACHECO J, JOSE R A, JUAN M R M. Visual simultaneous localization and mapping: a survey[J]. Artificial Intelligence Review, 2015, 43(1):55 - 81.

[7] DAVISON A J, REID I D, MOLTON N D, et al. MonoSLAM: real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6):1052 - 1067.

[8] KLEIN G, MURRAY D. PARALLEL Tracking and mapping for small AR workspaces[C]// IEEE & ACM International Symposium on Mixed & Augmented Reality. ACM, 2008: 225 - 234.

[9] ENGEL J, SCHPS T, CREMERS D. LSD - SLAM: large-scale direct monocular SLAM[C]// European Conference on Computer Vision. Springer, Cham, 2014.

[10] RAVL Mur_Artal, JUAN D Tordos. ORB - SLAM2: an open-source SLAM system for monocular, stereo and RGB - D cameras[J]. IEEE Transactions on Robotics, 2017,33(5):1 - 8.

[11] CALONDER M, LEPETIT V, STRECHA C, et al. BRIEF: binary robust independent elementary features[M]. Springer, Berlin, Heidelberg, 2010.

[12] MA T, WANG Y, WANG Z, et al. ASD - SLAM: a novel adaptive-scale descriptor learning for visual SLAM[C]// 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020.

[13] LI X, LIN S, XU M, et al. PO - SLAM: a novel monocular visual SLAM with points and objects[C]//2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), 2021.

[14] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: self-supervised interest point detection and description[C]// IEEE/CVF. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018: 00060.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 1409.1556,2014.

[16] TANG J, ERICSON L, FOLKESSON J, et al. GCNv2: Efficient correspondence prediction for real-time SLAM[J]. IEEE Robotics and Automation Letters, 2019,4(4):3505 - 3512.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770 - 778.

[18] HU F, CHENG J, BAO Y, et al. Accuracy enhancement for the front-end tracking algorithm of RGB - D SLAM[J]. Intelligent Service Robotics, 2020, 13(6):207 - 218.

[19] 刘强, 段富海, 桑勇, 等. 复杂环境下视觉 SLAM 闭环检测方法综述[J]. 机器人, 2019, 41(1):770 - 718.

[20] 袁朝春,王桐,何友国,等.传感器感知盲区条件下智能汽车主动制动系统控制研究[J]. 农业机械学报,2020,51(2):363 - 373.
YUAN Chaochun,WANG Tong,HE Youguo,et al. Intelligent vehicle active braking system control under sensor occluded scenes[J]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(2):363 - 373. (in Chinese)

[21] 徐春博,赵丁选,孔维天,等.基于视觉的吊装机器人卷扬随动控制研究[J]. 农业机械学报,2020,51(1):417 - 426.
XU Chunbo,ZHAO Dingxuan,KONG Weitian,et al. Visual based robotic hoisting motion following control[J]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(1):417 - 426. (in Chinese)

[22] PARK S, SCHOPS T, POLLEFEYS M. Illumination change robustness in direct visual SLAM[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017.