面向轻轨的高精度实时视觉定位方法

王婷娴 [1, 2, 3] 贾克斌 [1, 2, 3] 姚 萌 [1]

摘 要 轻轨作为城市公共交通系统的重要组成部分, 对其实现智能化的管理势在必行. 针对城市轻轨定位系统要求精度高、实时强且易于安装等特点, 本文提出一种基于全局−局部场景特征与关键帧检索的定位方法. 该方法在语义信息的指导下, 从单目相机获取的参考帧中提取区别性高的区域作为关键区域. 并结合像素点位置线索利用无监督学习的方式筛选关键区域中描述力强的像素对生成二值化特征提取模式, 不仅能够提升匹配精度还显著提高了在线模块场景特征提取与匹配的速度. 其次, 以场景显著性分数为依据获取的关键帧避免了具有相似外观的场景给定位带来的干扰, 并能辅助提高场景在线匹配的精度与效率. 本文使用公开测试数据集以及具有挑战性的轻轨数据集进行测试. 实验结果表明, 本系统在满足实时性要求的同时, 其定位准确率均可达到 90% 以上.

关键词 视觉定位, 位置识别, 关键帧检索, 关键区域检测, 序列匹配

引用格式 王婷娴, 贾克斌, 姚萌. 面向轻轨的高精度实时视觉定位方法. 自动化学报, 2021, **47**(9): 2194−2204

DOI 10.16383/j.aas.c200009

Real-time Visual Localization Method for Light-rail With High Accuracy

WANG Ting-Xian[1, 2, 3] JIA Ke-Bin[1, 2, 3] YAO Meng[1]

Abstract As an important part of the urban public transportation system, it is imperative to realize the intelligent management of light rail. By considering the practical requirements like high accuracy, real-time performance, and easy installation, this paper proposes a visual localization method based on global-local features and keyframe retrieval. Under the guidance of semantic information, the region with high significance in each reference frame obtained by the monocular camera is extracted as the key region. Combined with the location cues of pixels, unsupervised learning is used to filter the pixel pairs with strong description force in the key region to generate the binary pattern, which greatly reduces the computation of feature extraction and matching in the online module while improving the matching accuracy. Secondly, the keyframes obtained based on the discrimination score can effectively avoid the interference caused by the scene with analogous appearance, and assist to improve the accuracy and efficiency of online scene matching. The Nordland dataset and the challenging light rail dataset are used for testing. The experimental results show that the precision of the system can reach more than 90% while meeting real-time requirements.

Key words Visual localization, place recognition, keyframe retrieval, key region detection, sequence matching

Citation Wang Ting-Xian, Jia Ke-Bin, Yao Meng. Real-time visual localization method for light-rail with high accuracy. *Acta Automatica Sinica*, 2021, **47**(9): 2194−2204

轻轨在城市公共交通系统中扮演着重要的角色. 其因运行环境复杂, 难以实现信息化和智能化管理, 在一定程度上影响行车安全. 因此, 开发针对轻轨的高级驾驶辅助系统 (Advanced driver assistance system, ADAS) 势在必行[1]. 列车定位技术作为自动驾驶系统的重要组成部分之一, 能否提供准确地轻轨定位信息, 直接影响行车安全和调度的有效性. 目前查询应答器、里程计以及全球定位系统等技术被普遍地应用于列车实时定位任务中[2]. 在城市环境中, 由于建筑群、隧道的普遍存在, 通常会因遮挡产生多路径效应导致小范围内定位出现偏差[3]; 目前广泛应用的地面查询应答器属于非连续定位且需要昂贵的运营维护成本; 基于里程计的定位技术依赖于特殊的传感器, 比如车轮测速传感器、惯性测量传感器, 这两种传感器都会受到列车本身的影响从而随列车的运行产生累计误差[4−5]. 在此情

收稿日期 2020-01-08 录用日期 2020-05-28
Manuscript received January 8, 2020; accepted May 28, 2020
国家重点研发计划 (2018YFF01010100), 国家自然科学基金 (61672064), 青海省基础研究计划 (2020-ZJ-709) 资助
Supported by National Key Research and Developmet Program of China (2018YFF01010100), National Natural Science Foundation of China (61672064), and Basic Research Program of Qinghai Province (2020-ZJ-709)
本文责任编委 刘青山
Recommended by Associate Editor LIU Qing-Shan
1. 北京工业大学信息学部 北京 100124 2. 先进信息网络北京实验室 北京 100124 3. 计算智能与智能系统北京市重点实验室 北京工业大学 北京 100124
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Laboratory of Advanced Information Networks, Beijing 100124 3. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124

况下, 基于单目相机的定位方法由于其成本效益和信息的丰富性, 在轻轨定位系统中可以发挥关键作用[6]. 对于视觉定位系统而言, 其主要包含三个关键部分: 场景特征提取, 路径地图构建, 以及真实匹配生成三大模块[7−8]. 由此可见, 一个综合定位系统的复杂度会非常高, 想要使整个系统达到实时性, 更需要从整体考虑使各个模块算法的时间开销都尽可能地减少.

面对复杂的运行环境, 如何快速地从图像中提取鲁棒性好的特征描述符一直是视觉定位领域研究的热点. 近年来, 视觉定位任务中常用的特征提取方法主要有两类, 分别是手工制作特征[6, 9−12] 和深度学习特征[13−19]. 目前基于局部特征的视觉定位方法通常在光照、季节等因素导致环境外观出现明显变化时表现不佳[9−10]. 主要是因为基于特征点的描述符缺少整体的结构信息易导致感知混淆, 从而降低局部描述符的辨别力. 此外, 基于全局特征的方法虽然相较于前者具有更好的条件不变性, 但对视点变化和遮挡的鲁棒性较差[11]. 为弥补两者各自的缺点, 将局部和全局特征有效融合在一起是目前研究的趋势[6, 12]. 然而, 这类算法虽能够有效提升定位准确率却因特征向量维数的激增与场景规模的扩大无法满足实时性的要求.

近年来, 深度卷积神经网络在特征提取方面取得突出的成绩[13−14, 16]. 相比于 Gist、Fisher vector、VLAD 等手工制作特征, 深度学习特征在光照变化的环境下具有更好的识别能力[14, 16]. 因此越来越多的研究者为获得更精确的定位结果, 将卷积神经网络作为一种特征提取手段应用于视觉定位任务中[17−19]. 目前许多方法结合目标检测先提取显著性高的区域作为地标候选者, 再利用卷积神经网络描述该区域稳定特征, 最后筛选出潜在地标并将其映射到低维空间从而生成特征描述符. 这类方法因融合了全局和局部特征描述方法的优势, 在视觉定位领域卓有成效[17−18]. 此外, 还有的方法通过人工标定方式构建丰富的场景识别数据库, 并利用这些数据反复学习到稳定的场景特征, 虽然深度学习特征具有更好的稳定性, 但仅靠全局描述符作为场景匹配的依据仍难以实现鲁棒性高的定位, 且昂贵的时间成本和巨大的人工成本根本无法满足实际应用的需求.

针对上述问题, 本文设计了一种高精度实时视觉轻轨定位系统, 该系统的创新之处主要体现在以下 4 个方面:

1) 提出一种衡量像素显著性的方法来识别参考序列中的关键帧, 为在线模块提供适合的检索窗口, 避免因列车经停站造成的离散相似场景的影响, 同时有效提升了大规模复杂环境下场景匹配过程的计算效率.

2) 为消除场景中掺杂的不稳定信息对定位结果产生的干扰, 本文提出一种融合语义特征的关键区域检测方法, 在减少特征提取运算量的同时有效地保留了场景中有助于列车定位的显著性信息.

3) 提出一种无监督学习结合像素点位置线索的二值化特征描述方法, 在降低场景匹配计算复杂度的同时因不受描述区域形状的限制具有更广的应用范围. 在场景跟踪中, 将该描述符与场景序列匹配算法相结合能够克服因高帧率造成的连续相似场景使定位精度降低的问题.

4) 本系统只需单目相机采集数据既可, 兼容性高且可移植性强, 并实现了视觉定位任务对实时性和高精度的要求.

本文结构如下: 第 1 节概述了所提轻轨定位系统的构成以及详细地描述了系统中各个模块所涉及的关键技术; 第 2 节展示了实验结果并对结果进行分析讨论. 最后是结论部分.

1 面向轻轨的高精度实时视觉定位方法

1.1 系统框架

系统的整体框架如图 1 所示, 可概括为 4 个子模块: 1) 首先对预处理后的参考序列利用所提出的像素显著性计算方法得到每帧图像的显著性分数, 以其作为依据筛选出序列中的关键帧; 2) 其次利用语义分割网络生成的二值化掩模与像素显著性分数作为评判依据, 建立参考序列中每帧图像的关键区域; 3) 然后利用无监督学习方法并结合像素点位置信息得到场景特征抽取模式, 为后续在线模块快速生成二值化特征描述符做准备; 4) 最后, 通过离线部分获得的关键帧与场景特征提取模式完成在线匹配, 获取轻轨实时位置.

1.2 关键帧识别

轻轨在运行过程中, 经常会因为经停站造成不同位置场景内容离散相似的问题, 这增大了视觉定位任务的难度. 针对该问题, 本研究提出一种基于像素显著性分数的关键帧识别算法. 获取的关键帧为后续场景跟踪提供合适的检索窗口, 提升匹配精度.

在场景匹配中, 采集到的参考帧中通常会保留影响定位精度的不稳定信息, 因此需建立感兴趣区域 (Region of interest, ROI) 将参考帧中包含运动物体、铁轨和边缘模糊的区域移除. 衡量视频帧重要程度的显著性分数越高则表示该帧包含的特异性信息越多. 本方法利用滑动窗口遍历感兴趣区域中

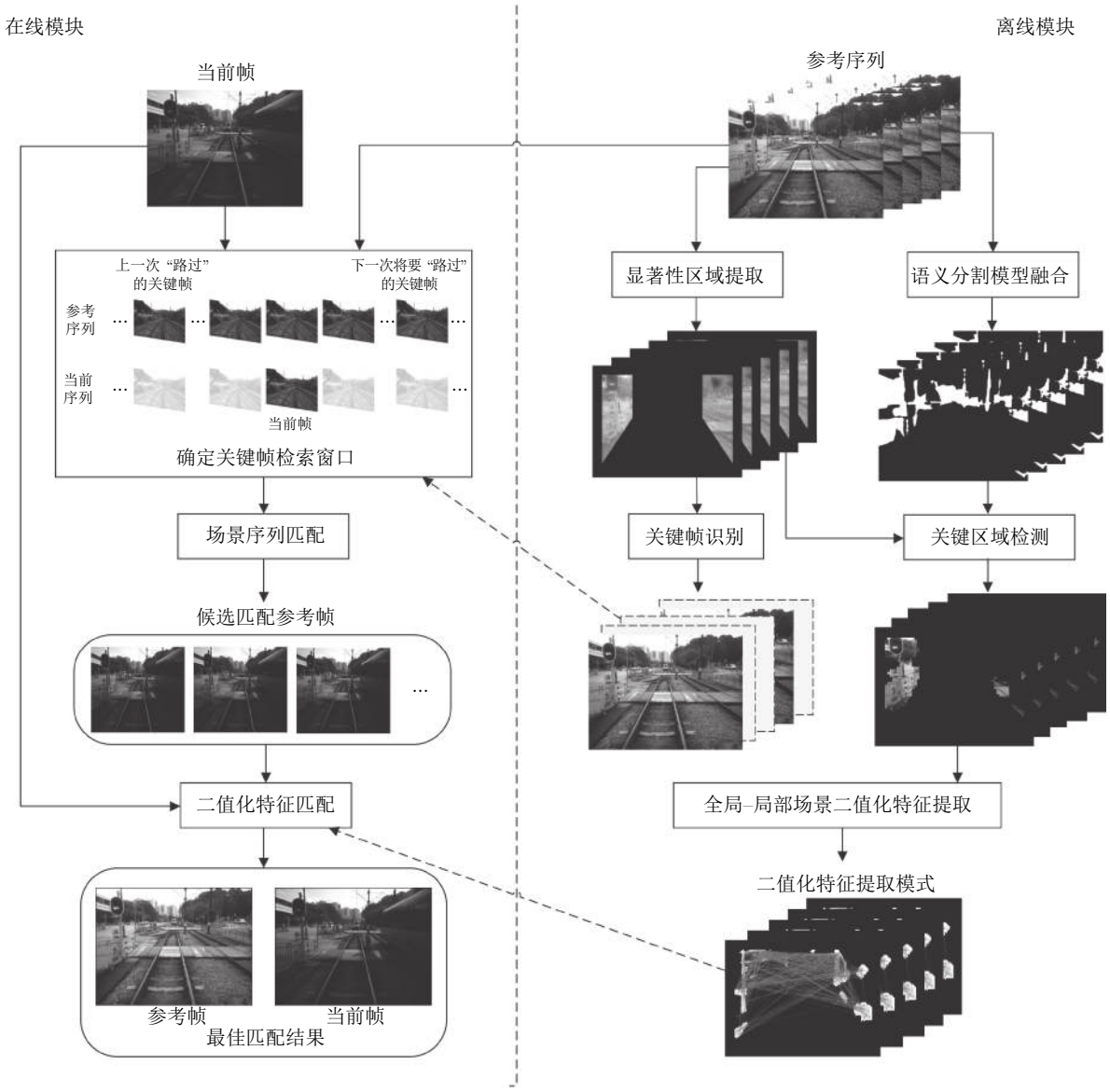


图 1 所提轻轨定位系统框架

Fig. 1 The framework of our proposed light-rail localization system

所有像素, 从而计算得到像素显著性分数.

记当前待计算的视频帧为 f_t, 其时域邻域内包含 N 个视频帧 (图 2 中以 $N=4$ 为例). 当滑动窗口处于像素点 (x,y) 位置时, 如式 (1) 所示, 分别计算当前帧所包含的图像块 $R(x,y,f_t)$ 与其他视频帧相同位置以及其十字邻域内, 5 个图像块 $R(x\pm1,\ y\pm1,\ f_t)$ 之间的差别, 使求和得到的当前帧 (x,y) 位置处像素分数既在时间域和空间域具有显著性又均衡因外界不可抗力因素造成的抖动影响.

$$S_p(x,y,f_t)=\frac{1}{5N}\sum_{f_{t'}\in N,t'\neq t} D_E(R(x,y,f_t),\ R(x\pm1,y\pm1,f_{t'})) \quad (1)$$

其中, $D_E(\cdot)$ 表示图像块间的差别, 通过方向梯度直方图 (Histogram of oriented gradient, HOG) 特征利用欧氏距离计算得到, 以降低光线变化带来的干扰. $R(x\pm1,\ y\pm1,\ f_t)$ 是序列中其他帧相同位置及十字邻域内的图像块. $S_p(x,y,f_t)$ 表示像素的显著性分数.

将 ROI 中所有像素显著性分数求和得到视频帧的显著性分数. 关键帧作为划分参考序列的标志, 其显著性分数应在全局和局部范围内都高于一般的视频帧, 因此利用关键帧检索窗口能够在场景匹配中获得高置信度的匹配结果. 关键帧提取主要分为两步, 先基于适当的检索窗口提取参考序列局部范围内显著性分数最高的视频帧, 并对显著性分数进行降序排序; 其次将前 N_k 帧作为关键帧. 如图 3 所

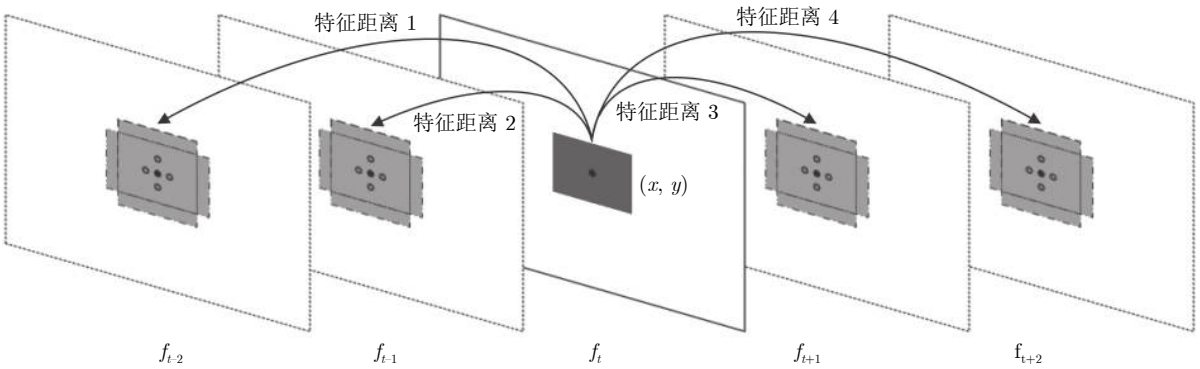


图 2　计算像素显著性分数的示意图

Fig. 2　Illustration of computing saliency score of pixel

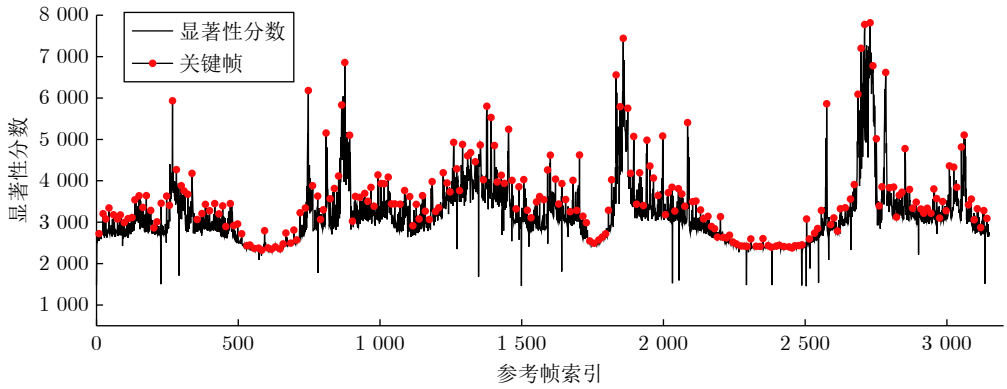


图 3　关键帧在参考序列中的分布

Fig. 3　Distribution of keyframes in the reference sequence

示, 这些关键帧均匀分布在整个参考序列, 为后续场景匹配提供了稳定的跟踪锁定.

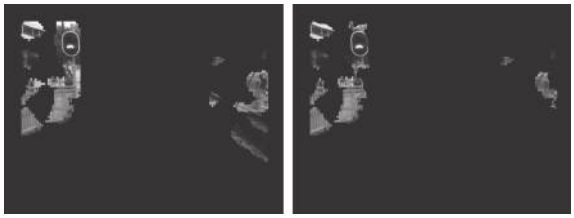
1.3　关键区域检测

针对视觉定位任务, 出于高精度和低计算复杂度兼得的考量, 本文提出一种融合语义特征的关键区域检测方法. 关键区域是参考帧中包含特异性信息的区域, 这些信息不随时间的变化而变化, 能有助于提高定位结果的鲁棒性. 期望检测到的关键区域中尽可能少地包含背景信息和动态目标, 类似天空、树木、车辆等. 因为这些场景在序列中普遍存在, 定位时不能提供有效的信息.

为了减少冗余信息对场景匹配的干扰, 同时提高后续提取二值化特征的效率. 所提出的关键区域检测算法步骤如下: 首先利用之前计算像素显著性分数的方法, 将显著性分数高于一定阈值 T_K的像素保留作为初步关键区域, 检测结果如图 4(a) 所示. 因每帧中像素显著性分数分布在不同尺度, 故而采用自适应阈值 T_K, 从而保障算法的鲁棒性. 阈值计算方式如式 (2) 所示

$$T_K(f_t) = K \times \frac{1}{N_p} \sum_{(x,y)\in ROI} S_p(x,y,f_t) \qquad (2)$$

其中, N_p 为 ROI 中像素的总个数, K 为间接调整阈值的系数. 从图 4(a) 中发现, 提取到的特征区域虽然保留住图像中特异性场景, 但是仍混入了无用信息. 针对该问题, 本文使用多个在 Cityscapes 数据集上训练的网络模型对参考帧进行语义分割. 按照特异性和稳定性的原则, 分割时只保留所需要的 6 类场景分别是: 建筑物、墙、电线杆、围栏、信号灯、标志牌. 对不同模型分割后的结果, 再通过加权融



(a) 初步关键区域
(a) Rough key region

(b) 精细化关键区域
(b) Refined key region

图 4　关键区域检测结果

Fig. 4　The result of key region detection

合的方式生成分割精度更高的二值化掩膜. 最后将前两步检测到的特征区域取交集, 得到精细化后的关键区域 (如图 4(b) 所示).

1.4 全局–局部场景二值化特征提取

场景匹配模块作为实时处理单元, 不仅要求场景特征描述符具有高区分性还需降低计算特征间相似度的复杂度. 基于之前获取的关键区域, 使用 HOG 或尺度不变特征变换 (Scale-invariant feature transform, SIFT) 等特征描述符进行场景匹配, 虽然可以获得更精确的匹配结果, 但巨大的计算复杂度难以满足系统实时性的要求. 二值化特征描述符因使用汉明距离计算特征间相似度, 能够大幅度提升匹配效率. 受此启发, 本文提出一种基于无监督学习的全局–局部场景二值化特征提取方法. 该方法利用一种新颖的像素对筛选机制, 使保留下来的像素对包含丰富的空间和上下文信息.

当前帧的特征描述符是由级联所筛选出像素对的二值化比较结果得到的. 只有提取到描述力强的像素对才能增加描述符的区分力. 本文利用式 (3), 计算得到像素对显著性分数, 由此来评估其辨别度.

$$S_{\text{pair}}(P,f_t)=\sum_{i=0}^{N_f}\left(D_I(P,f_i)-D_I(P,f_t)\right) \tag{3}$$

其中, $S_{\text{pair}}(P,f_t)$ 是当前帧 f_t 内某点对 P 的显著性分数, $D_I(P,f_t)$ 是当前帧 f_t 内点对 P 的两像素间的灰度差, $D_I(P,f_i)$ 是第 i 个相邻帧内点对 P 的两像素间的灰度差. N_f 是相邻帧的数量.

除此之外, 提取到的所有像素对还需要包含丰富的空间信息. 在关键区域中通常包含两类像素对. 一种是两个像素来自相同特征子区域; 另一种是像素来自不同特征子区域. 两者二值化的结果分别保留了图像中的局部细节信息和全局结构信息. 然而, 根据关键区域获取的所有像素对, 因数量庞大, 即使离线处理仍会耗费巨大的计算内存和时间成本. 因此, 本文引入像素点位置线索对区域内包含的像素对进行初步筛选. 将具有相同位置线索的像素对按照显著性分数降序排列, 只保留其中排名靠前的 N_t 对像素点作为初步筛选结果.

像素对来自不同区域, 会包含不同的信息. 保留空间相关性高的像素对会使描述符的区分力降低, 因此本文基于原型聚类算法进一步筛选得到相关性低的点对. 首先, 基于像素对的初步筛选结果, 利用式 (4) 逐一计算其分布向量构建对于当前帧的训练样本集 D.

$$\boldsymbol{x}_1(P_1,f_t)=\begin{pmatrix}\Delta(P_1,f_{t-m})\\ \Delta(P_1,f_{t-m+1})\\ \vdots\\ \Delta(P_1,f_t)\\ \vdots\\ \Delta(P_1,f_{t+m})\end{pmatrix} \tag{4}$$

$$\Delta(P,f_t)=\Delta P(p_i,p_j,f_t)=I(x_i,y_i,f_t)-I(x_j,y_j,f_t) \tag{5}$$

其中, 分布向量 $\boldsymbol{x}_1(P_1,f_t)$ 表示像素对 P_1 中所对应像素 p_i 和 p_j 间的灰度值之差在视频帧 f_i 中的分布, $i\in[t-m,\ t+m]$. $I(\cdot)$表示像素的灰度值, 基无监督学习的场景特征模式提取过程伪代码如算法 1 所示.

算法 1. 基于无监督学习的场景特征模式提取过程伪代码

输入: 训练样本集 $D=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{N_t}]$; 预定义筛选像素对个数 N_c.

过程:

1: 从 D 中随机选取 1 个样本作为初始聚类中心 $\boldsymbol{\mu}_1$

2: repeat

3: for $j=1,2,\cdots,N_t$ **do**

4: 计算分布向量 $\boldsymbol{x}_j$ 与初始聚类中心 $\boldsymbol{\mu}_1$ 的距离为

$$d_j=\|\boldsymbol{x}_j-\boldsymbol{\mu}_1\|_2$$

5: end for

6: for $j=1,2,\cdots,N_t$ **do**

7: 计算分布向量 $\boldsymbol{x}_j$ 被选为下一个聚类中心的概率为

$$p_j=\frac{d_j^2}{\sum\limits_{j=1}^{N_t}d_j^2}$$

8: end for

9: 选取概率最大值对应的样本作为下一个聚类中心 $\boldsymbol{\mu}_2$

10: until 选择出 N_c 个聚类中心

11: repeat

12: 令 C_i 为空集, $1\le i\le N_c$

13: for $j=1,2,\cdots,N_t$ **do**

14: 计算分布向量 $\boldsymbol{x}_j$ 与聚类中心 $\boldsymbol{\mu}_i$ 的距离为

$$d_{j,i}=\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|_2$$

15: 以距离为依据确定 $\boldsymbol{x}_j$ 所属簇 C_{τ_j}

$$\tau_j=\mathop{\arg\min}_{i\in\{1,2,\cdots,N_c\}}d_{j,i}$$

16: end for

17: for $i=1,2,\cdots,N_c$ **do**

18: 依据所划分的簇 C_i, 计算新聚类中心

$$\boldsymbol{\mu}_i'=\frac{1}{|C_i|}\sum_{x\in C_i}\boldsymbol{x}$$

19: if $\boldsymbol{\mu}_i' \neq \boldsymbol{\mu}_i$ **then**

20: 更新当前聚类中心为 $\boldsymbol{\mu}_i'$

21: else

22: 聚类中心保持不变

23: end if

24: end for

25: until 聚类中心不再变化

26: 遍历所有簇 C_i, 将簇中与聚类中心 $\boldsymbol{\mu}_i$ 距离最近的分布向量 $\boldsymbol{y}_i$ 作为最终的结果.

输出: 结果集 $R = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{Nc}]$.

1.5　在线场景序列匹配

图像序列匹配算法[20−22] 因结合了时域信息和序列图像的一致性, 即使在环境外观复杂变化的情况下, 获得的匹配结果也具有较高的鲁棒性. 其中, 最具代表性的就是 SeqSLAM 算法[20], 但该算法对因高帧率造成的连续相似场景和因轻轨经停站造成的离散相似场景区分力不足. 针对此缺陷, 本文利用第 1.2 节中获取的关键帧, 将参考序列划分为多个场景间区分度大的子序列作为当前帧的候选检索窗口, 在提高场景匹配效率的同时有效地控制了定位误差的范围.

如图 5 所示, 上一个匹配过的关键帧 f_{last} 到下一个将要匹配的关键帧 f_{next} 之间的范围作为当前帧 f_t 的检索窗口. 直接使用所提方法在离线部分生成的场景特征提取模式, 获得 f_t 与 f_{next} 的二值化特征向量, 并计算两者之间的汉明距离 $D_H(f_t,\ f_{\text{next}})$. 根据该距离与阈值 T_L 间的大小关系, 从而确定检索窗口 E_t 的范围. 若 $D_H(f_t,\ f_{\text{next}}) \leq T_L$, 则 f_t 对应的检索窗口将移动到下一个场景子序列, 即令 f_{next} 作为新的 f_{lsat}, 将关键帧集合中居于 f_{next} 之后的相邻关键帧作为新的 f_{next}; 反之, 则 f_t 对应的检索窗口保持不变.

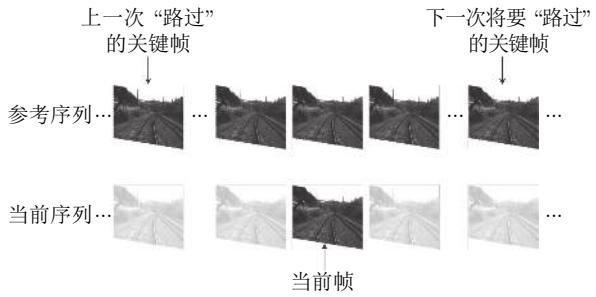


图 5　在线场景序列匹配中的关键帧检索窗口

Fig. 5　Illustration of the keyframe retrieval window for online sequence

利用场景序列匹配算法在 f_t 对应的检索窗口 E_t 内搜索与 f_t 最匹配的参考帧并建立候选匹配参考帧集合 Q_t. 然后, 对属于 Q_t 的任意参考帧 f_i, 通过特征模式计算其对应的描述符, 记为 $B_i(f_i)$; 使用同一特征模式计算当前帧 f_t 的场景描述符, 记为 $B_i(f_t)$. 通过式 (5), 可检索到 Q_t 内与 f_t 最佳匹配的参考帧 f_{matched}.

$$f_{\text{matched}}(f_t) = \arg \min_{f_i \in Q_t} (D_H(B_i(f_t), B_i(f_i))) \tag{6}$$

2　实验与分析

2.1　数据集

实验中使用的 MTRHK 数据集和 Nordland 数据集分别由中国香港港铁 (Mass Transit Railway, MTR) 和挪威广播公司 (Norwegian Broadcasting Corporation, NRK) 提供[6]. MTRHK 数据集采集自轻轨 507 号路线, 包含 3 组视频序列, 共 13 859 帧. 分辨率为 640×480 像素, 帧率为 25 帧/s, 每组视频序列包含两段序列, 其是从同一列车在相同的路径上不同运行时间采集到的[5]. 由于采集时间不同, 序列间存在环境以及列车速度变化, 需人工校准作为真实标定. 此外, 该数据集中包含了大量具有挑战性的场景, 如图 6(a)~6(c) 所示. Nordland 数据集包含 4 个季节采集的视频序列, 原始分辨率为 1 920×1 080 像素, 帧率为 25 帧/s, 其场景包含城市以及自然等不同类型环境[12]. 本文选取秋季和夏季共 12 000 帧作为训练和测试数据, 并降采样至 640×480 像素. 这两段序列采集自相同运行速度, 故而具有相同帧号的视频帧采集自相同的位置.

2.2　评价方法

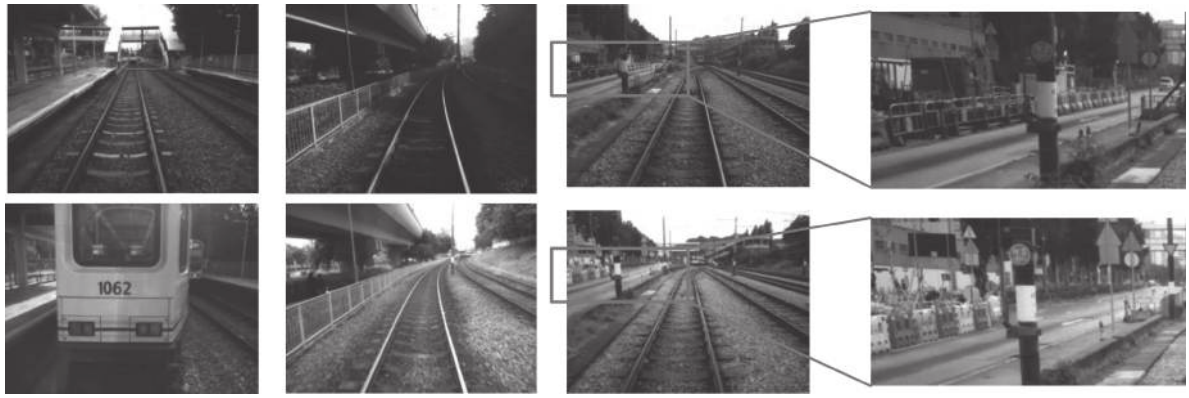
本文将轻轨定位任务近似作图像检索任务, 将准确率和召回率用来评价所提出方法的性能. 对于 Nordland 数据集, 由于列车运行速度保持不变, 与当前帧具有相同帧号的参考帧可直接作为真实标定; 对于 MTRHK 数据集, 手动标定不同序列间视频帧的对应关系, 将此结果作为真实标定. 在实验中, 将匹配结果与真实标定间的差别, 称为匹配偏差, 单位为帧. 若匹配上的两幅场景称为阳性样本, 则匹配偏差大于容差范围的阳性样本称为假阳性样本 (False positives, FP), 反之称为真阳性样本 (True positives, TP).

2.3　参数设置

在实验中, 所提出方法针对不同数据集所需参数的默认值存在差异, 如表 1 所示.

2.4　单帧场景识别

通过分割模型得到的语义特征对关键区域检测



(a) 车辆遮挡
(a) Vehicle occlusion

(b) 光照变化
(b) Illumination changes

(c) 场景内容变化
(c) Partly scene content changes

图 6 中国香港轻轨数据集中复杂多变的场景示例

Fig. 6 Examples of complex and volatile scenes in the China Hong Kong light-rail dataset

表 1 Nordland 和 MTRHK 数据集中所需参数设置

Table 1 Parameter settings for Norland and MTRHK datasets

参数符号	参数定义	参数值 (Nordland)	参数值 (MTRHK)
$V_{\min}$	最小路径拟合速度	0.8	0
$V_{\max}$	最大路径拟合速度	1.2	1.5
V_{step}	路径拟合速度步长	0.1	0.1
N_c	像素对提取个数	512	512
K	关键区域检测系数	1.05	1.05
T_L	最佳匹配距离阈值	175	175

具有指导作用. 由于单一模型的性能存在局限性, 本文通过模型融合的方式将不同分割网络获得的语义信息有机地结合在一起, 从而优化最终的分割效果. 本文从参考序列中筛选出 50 个关键帧进行人工标定, 用标定真值与分割结果计算平均交并比. 由表 2 结果可知, 融合后得到的分割效果明显优于单个模型. 对于场景更为复杂的轻轨数据集而言, 效果提升尤为明显. 图 7 对分割结果进行了可视化展示.

表 2 不同语义分割模型间的精度对比

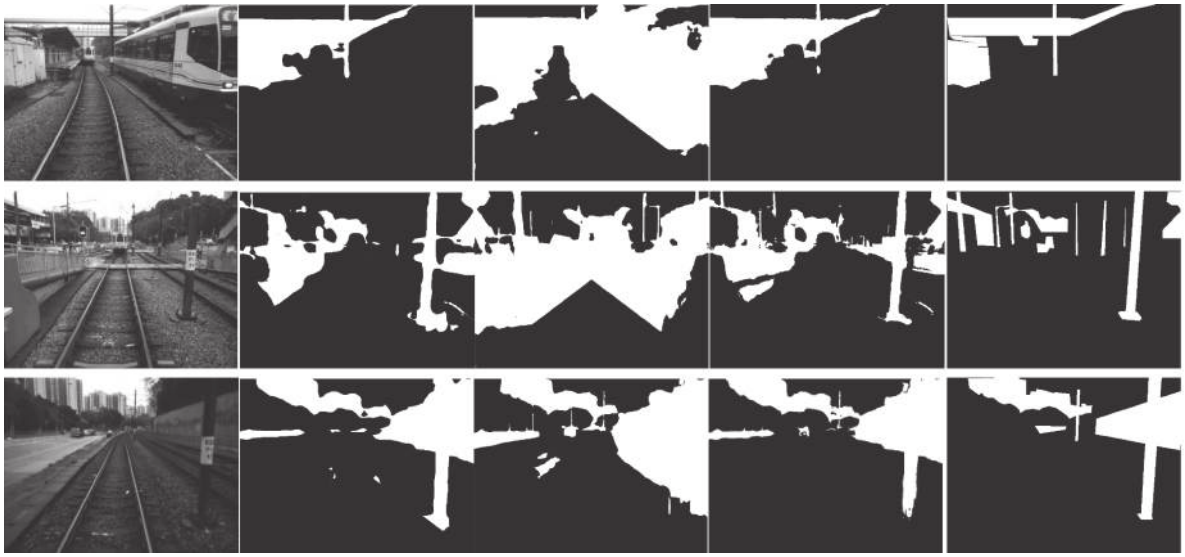
Table 2 Accuracy comparison of different semantic segmentation network

语义分割网络	平均交并比 (%)	
	Nordland	MTRHK
FCN	67.9	54.9
PSPNet	70.8	32.7
Deeplab	71.7	55.8
RefineNet	72.5	59.2
DFN	73.0	48.2
BiSeNet	72.2	36.2
融合模型	78.0	64.6

为来验证关键区域检测方法的有效性, 实验中利用 HOG 特征作为描述符, 比较了特征描述区域大小不同的五种场景描述方法. 通过匹配偏差反映场景识别的质量, 匹配偏差越小则表示匹配效果越好. 方法 1 将整个视频帧作为特征描述区域计算一个 HOG 特征; 方法 2 将整个视频帧分割成 40×40 互不重叠的图像块, 分别计算 HOG 特征. 匹配时, 计算两幅图像对应位置小块的 HOG 特征向量间的欧氏距离, 并将所有的欧氏距离相加得到图像间的相似度; 方法 3 与方法 2 类似, 匹配时只考虑 ROI 中包含的图像块; 方法 4 是对基于像素显著性分数检测到的特征区域进行描述, 其场景描述符通过计算每个连通的关键区域的 HOG 特征获得. 方法 5 是在方法 4 的基础上融合语义信息获取关键区域, 再进行描述符的提取.

图 8 显示了这五种方法的匹配偏差和时间成本. 纵轴为匹配偏差, 横轴为计算时间以对数刻度方式呈现. 如图 8 所示, 全局 HOG 特征方法的匹配偏差最高, 局部 HOG 特征方法的计算时间最长. 相比于前两者, 基于特征区域的局部 HOG 特征方法有效地权衡了计算效率与匹配质量间关系. 虽然与局部 HOG 特征相比, 基于感兴趣区域的场景匹配方法因描述区域的缩小导致匹配精度下降但是计算效率有显著提升. 此外, 融合语义信息后所得到的匹配精度最高. 由此可见, 在计算场景特征描述符前, 先对图像进行关键区域检测是必不可少的. 不仅能够减少无用信息的干扰提升匹配精确度, 还能大幅度缩减时间成本.

如前文所述, 所提出的关键区域检测方法需要根据像素显著性分数和自适应阈值 T_K 对视频帧提取初步特征区域. 根据式 (2) 可知 T_K 是帧内平均



(a) 原始图像 (a) Original image　(b) Deeplab 模型 (b) Model of Deeplab　(c) BiSeNet 模型 (c) Model of BiSeNet　(d) 融合模型 (d) Model fusion　(e) 真实标定 (e) Ground truth

图 7　不同语义分割网络获得的结果示例

Fig. 7　Example results of different semantic segmentation network

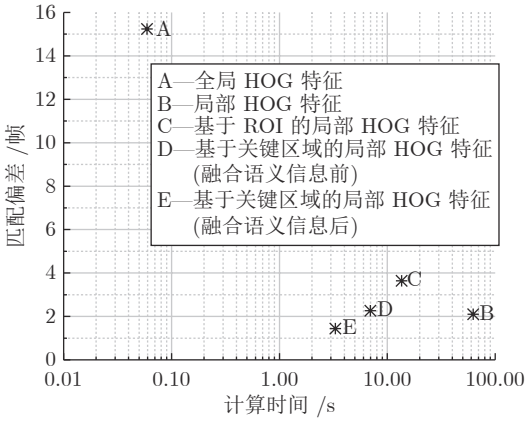


图 8　不同方法的匹配偏差和计算时间

Fig. 8　Matching offset and computation time of different methods

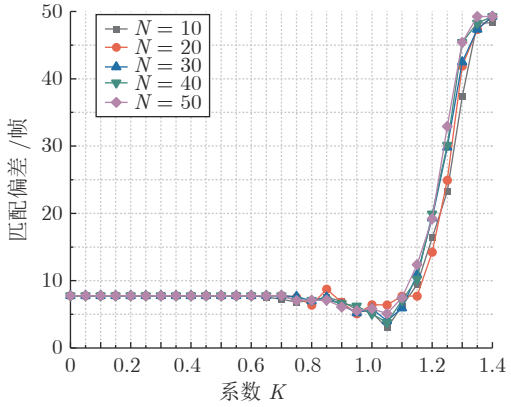


图 9　系数 K 对匹配精度的影响

Fig. 9　The influence of coefficient K for matching accuracy

显著性分数与系数 K 的乘积, 通过系数 K 能够间接调整阈值大小. 为确定系数 K 的值, 对所有可能的系数 K 通过改变式 (1) 中 N 的值 (变化范围 10~50), 获取 5 组不同的关键区域进行对比实验. 如图 9 所示, 横轴为系数 K 的值, 纵轴为匹配偏差, 图中不同线型分别代表不同取值的 N. 对比发现, 当系数 $K = 1.05$ 时, 其匹配偏差最低.

SeqSLAM 算法中使用归一化降采样图像作为全局特征描述符[20], 其因运算速度快常被用于实时场景匹配模块. 为验证所提特征提取算法的性能, 在实验中将其与 SeqSLAM 算法以及在轻轨定位方面做出突出贡献的基于无监督学习的轻轨实时定位 (Unsupervised learning-based localization for light-rail real time, LRT-ULL) 算法[6] 进行了对比, 观察这三种方法在单帧场景识别中的表现. 图 10 中只展示了三种方法在 4 组场景真实标定帧中的表现, 横轴为邻近帧与真实标定间的相对索引, 左侧纵轴为匹配距离, 右侧纵轴为匹配分数. 其中, 仅 SeqSLAM 算法用匹配距离来衡量. 匹配距离越小则代表场景越相似, 匹配分数越大则代表场景匹配程度越高.

SeqSLAM 算法的匹配结果显示, 真实标定附近大约 10 个参考帧均与当前帧的匹配距离为 0. 这表明基于全局特征的场景匹配方法无法区分连续相似场景. 通过观察图 11 匹配分数曲线可知, 基于全

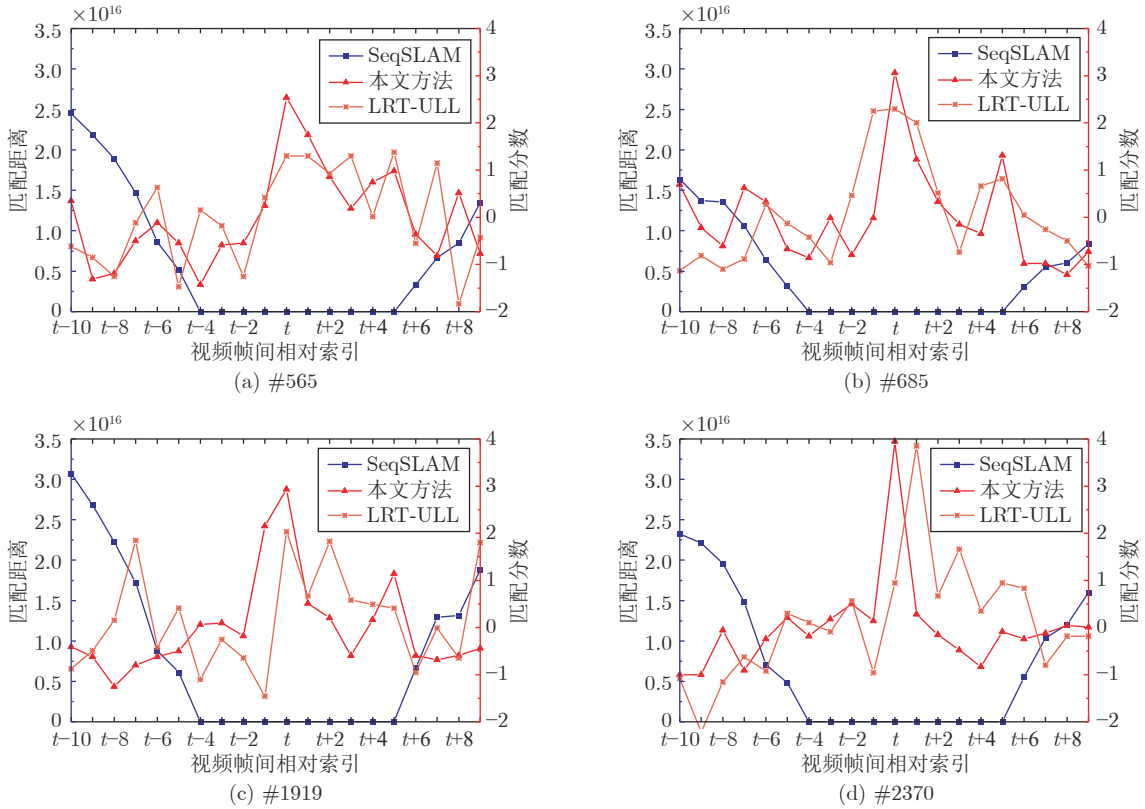


图 10 不同方法在单帧场景识别中的性能表现

Fig. 10 Performance of different methods in single frame scene recognition

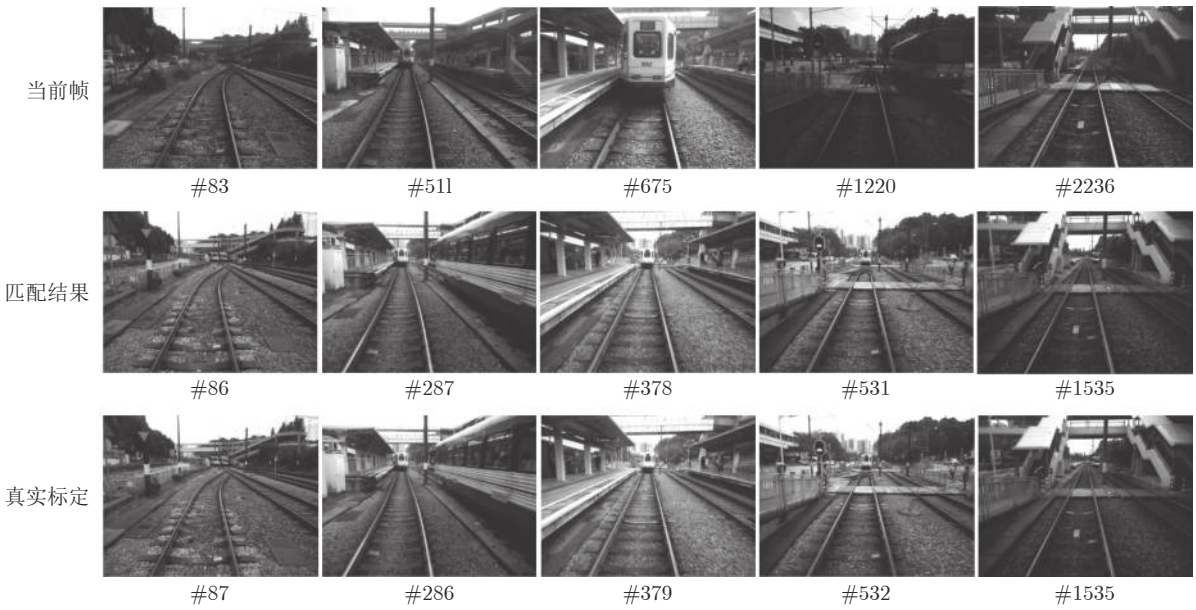


图 11 本文方法在 MTRHK 数据集中的匹配结果

Fig. 11 Illustration of matching results from the MTRHK dataset

局−局部特征的 LRT-ULL 算法与本文所提算法, 均能区分出连续场景间存在的差异. 但是, 所提算法匹配分数的峰值总是出现在真实标定位置, 而 LRT-ULL 算法存在较为明显的匹配偏差. 由此可见, 所提算法能够保留识别度高的特征, 对相似度高的连续场景具备显著的区分力, 能够对最终获得

精确的定位结果起到积极作用.

除此之外, 这三种方法对每帧图像的平均处理时间如表 3 所示. 所提出的场景特征提取算法明显比 LRT-ULL 算法速度快. 这是因为本文方法通过引入像素位置线索, 降低了训练矩阵的维数, 使场景描述效率得到显著提升.

表 3　不同方法对每帧图像的平均描述时间对比 (s)
Table 3　Comparison of average describing time for each image by different methods (s)

方法	SeqSLAM	LRT-ULL	本文方法
时间	0.1327	1.2791	0.1238

2.5　多帧场景跟踪

本文提出的定位方法通过引入关键帧检索机制, 避免了因长距离行驶和部分极端场景给匹配性能带来的影响. 表 4 为本文方法与 SeqSLAM 方法[20] 和 SeqCNNSLAM 方法[18] 分别在 Nordland 和 MTRHK 数据集上进行对比实验的结果. 可以看出, 在相同的的容差范围内, 针对场景复杂的 MTRHK 数据集, 本文所提出的定位方法在召回率为 100% 时精确度能达到 90.2%, 明显高于另外两种方法. 图 11 展示了在 MTRHK 数据集中, 所提出算法的场景匹配结果以及其真实标定. 对场景变化相对简单的 Nordland 数据集而言, 虽然本文方法在性能上稍逊于使用卷积神经网络提取场景特征的 SeqCNNSLAM 方法[18], 但精确度仍然能够到达 99.24%.

表 4　不同场景跟踪算法的准确率 (%) 与召回率 (%)
Table 4　Precision (%) and recall (%) of different scene tracking methods

数据集	准确率 (召回率)		
	SeqSLAM	SeqCNNSLAM	本文方法
Nordland	89.56 (100)	99.67 (100)	99.24 (100)
MTRHK	39.71 (100)	60.72 (100)	90.20 (100)

表 5 中对比了在 Nordland 数据集和 MTRHK 数据集中, 不同算法完成场景匹配时每帧的平均处理消耗时间. 该结果表明, 相比于全局特征和 CNN 特征, 使用本文方法计算得到的场景特征, 在场景匹配时所消耗时间最少. 这是因为该描述符使用汉明距离进行特征间相似度的计算, 大幅度提高了特征匹配的效率, 从而满足定位系统对实时性的要求. 结合表 4 和表 5 的实验结果可知, 本文所提出的定位方法实现了在匹配精度和计算复杂度之间的最佳平衡.

表 5　在 Nordland 数据集和 MTRHK 数据集中不同场景跟踪算法的消耗时间 (s)
Table 5　The consumption time of different scene tracking methods in the Nordland and the MTRHK dataset (s)

数据集	平均消耗时间		
	SeqSLAM	SeqCNNSLAM	本文方法
Nordland	0.67×10^{-1}	6.51×10^{-3}	3.17×10^{-3}
MTRHK	0.50×10^{-1}	4.90×10^{-3}	2.37×10^{-3}

3　结论

本文以高精度轻轨实时定位系统为研究背景, 采用视觉定位技术, 针对场景变化繁杂的轻轨运行环境, 提出结合关键帧检索机制和全局−局部场景二值化特征的定位方法. 该方法中, 通过融合语义特征检测到的关键区域既能有效降低计算时间成本, 又能提升场景识别的准确度. 其次, 在聚类算法的基础上结合像素位置线索筛选出低相关性的像素对, 不仅使提取到的场景描述符包含丰富的空间和上下文信息, 还减少了其中的冗余信息. 最后, 设计并实现了一个基于单目视觉信息的高精度轻轨实时定位系统. 实验结果表明, 本系统不仅解决了因高帧率造成的连续相似场景使定位精度降低的问题, 同时在场景内在结构发生变化等极端情况的干扰下依旧保持了较高的匹配精度, 既满足了轻轨定位系统对精度的要求也保证了实时性.

References

1　Martinez C M, Heucke M, Wang F-Y, Gao B, Cao D P. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2018, **19**(3): 666−676

2　Yu Yu-Feng, Zhao Hui-Jing, Cui Jin-Shi, Zha Hong-Bin. Road structural feature based monocular visual localization for intelligent vehicle. *Acta Automatica Sinica*, 2017, **43**(5): 725−734
(俞毓锋, 赵卉菁, 崔锦实, 查红彬. 基于道路结构特征的智能车单目视觉定位. 自动化学报, 2017, **43**(5): 725−734)

3　Bresson G, Alsayed Z, Yu L, Glaser S. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2017, **2**(3): 194−220

4　Ding Wen-Dong, Xu De, Liu Xi-Long, Zhang Da-Peng, Chen Tian. Review on visual odometry for mobile robots. *Acta Automatica Sinica*, 2018, **44**(3): 385−400
(丁文东, 徐德, 刘希龙, 张大朋, 陈天. 移动机器人视觉里程计综述. 自动化学报, 2018, **44**(3): 385−400)

5　Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 2016, **32**(6): 1309−1332

6　Yao Meng, Jia Ke-Bin, Siu Wan-Chi. Learning-based localization with monocular camera for light-rail system. *Journal of Electronics and Information Technology*, 2018, **40**(9): 2127−2134
(姚萌, 贾克斌, 萧允治. 基于单目视频和无监督学习的轻轨定位方法. 电子与信息学报, 2018, **40**(9): 2127−2134)

7 Piasco N, Sidibé D, Demonceaux C, Gouet-Brunet V. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 2018, **74**: 90−109

8 Lowry S, Sünderhauf N, Newman P, Leonard J J, Cox D, Corke P, et al. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 2016, **32**(1): 1−19

9 Cummins M, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, 2011, **30**(9): 1100−1123

10 Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 2015, **31**(5): 1147−1163

11 Naseer T, Burgard W, Stachniss C. Robust visual localization across seasons. *IEEE Transactions on Robotics*, 2018, **34**(2): 289−302

12 Qiao Y L, Cappelle C, Ruichek Y. Visual localization across seasons using sequence matching based on multi-feature combination. *Sensors*, 2017, **17**(11): 2442

13 Zhang X M, Zhao Z H, Zhang H J, Wang S Z, Li Z J. Unsupervised geographically discriminative feature learning for landmark tagging. *Knowledge-Based Systems*, 2018, **149**: 143−154

14 Hou Y, Zhang H, Zhou S L. Convolutional neural network-based image representation for visual loop closure detection. In: Proceedings of the 2015 IEEE International Conference on Information and Automation. Lijiang, China: IEEE, 2015. 2238−2245

15 Liu Li, Zhao Ling-Jun, Guo Cheng-Yu, Wang Liang, Tang Jun. Texture classification: State-of-the-art methods and prospects. *Acta Automatica Sinica*, 2018, **44**(4): 584−607
(刘丽, 赵凌君, 郭承玉, 王亮, 汤俊. 图像纹理分类方法研究进展和展望. 自动化学报, 2018, **44**(4): 584−607)

16 Sünderhauf N, Shirazi S, Dayoub F, Upcroft B, Milford M. On the performance of ConvNet features for place recognition. In: Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany: IEEE, 2015. 4297−4304

17 Kong Y G, Liu W, Chen Z P. Robust ConvNet landmark-based visual place recognition by optimizing landmark matching. *IEEE Access*, 2019, **7**: 30754−30767

18 Bai D D, Wang C Q, Zhang B, Yi X D, Yang X J. Sequence searching with CNN features for robust and fast visual place recognition. *Computers and Graphics*, 2018, **70**: 270−280

19 Arroyo R, Alcantarilla P F, Bergasa L M, Romera E. Fusion and binarization of CNN features for robust topological localization across seasons. In: Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, Korea (South): IEEE, 2016. 4656−4663

20 Milford M J, Wyeth G F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: Proceedings of the 2012 IEEE International Conference on Robotics and Automation. Saint Paul, MN, USA: IEEE, 2012. 1643−1649

21 Milford M. Vision-based place recognition: How low can you go? *The International Journal of Robotics Research*, 2013, **32**(7): 766−789

22 Pepperell E, Corke P I, Milford M J. All-environment visual place recognition with SMART. In: Proceedings of the 2014 IEEE International Conference on Robotics and Automation. Hong Kong, China: IEEE, 2014. 1612−1618

王婷娴 北京工业大学信息学部硕士研究生. 主要研究方向为图像处理与视觉信息定位.
E-mail: wangtingxian@emails.bjut.edu.cn
(**WANG Ting-Xian** Master student at the Faculty of Information Technology, Beijing University of Technology. Her research interest covers image processing and visual information localization.)

贾克斌 博士, 北京工业大学信息学部教授. 主要研究方向为图像/视频信号与信息处理. 本文通信作者.
E-mail: kebinj@bjut.edu.cn
(**JIA Ke-Bin** Ph.D., professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers image/video signal and information processing. Corresponding author of this paper.)

姚 萌 工程师. 2018 年获得北京工业大学信息学部博士学位. 主要研究方向为图像处理与视觉信息定位.
E-mail: yaomeng@emails.bjut.edu.cn
(**YAO Meng** Received his Ph.D. degree from the Faculty of Information Technology, Beijing University of Technology in 2018. His research interest covers image processing and visual information localization.)