基于单目视频和无监督学习的轻轨定位方法

姚 萌[①②] 贾克斌*[①③④] 萧允治[②]

[①](北京工业大学信息学部 北京 100124)

[②](香港理工大学电子讯息工程系 香港)

[③](先进信息网络北京实验室 北京 100124)

[④](未来网络科技高精尖创新中心 北京 100124)

摘 要：基于视觉信息的场景识别定位模块被广泛应用于车辆安全系统。针对目前场景逐帧匹配算法训练数据量大、匹配处理计算复杂度高以及跟踪精度低导致难以实际应用的问题，该文提出一种新的基于局部关键区域与关键帧的场景识别方法，在保证匹配精度的同时满足系统实时性的要求。首先，该方法仅使用单目摄像机捕获的单一序列作为参考序列，采用无监督方式提取序列的显著性区域作为关键区域，并计算关键区域中低相关性的二值化特征，提高了场景匹配的精确度并大幅减少了实时场景匹配过程中特征生成与匹配的计算复杂度。其次，该方法以显著性分数为依据提取参考序列中的关键帧，缩小了跟踪模块的检索范围并提高了检索效率。该文使用香港轻轨系统数据集以及公开测试数据集进行方法测试。实验结果表明，该文方法在实现快速匹配的同时，其匹配正确率较基于全局特征匹配方法SeqSLAM提高了9.8%。

关键词：视觉定位；关键区域；关键帧；二值化特征

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2018)09-2127-08

DOI: 10.11999/JEIT171017

Learning-based Localization with Monocular Camera for Light-rail System

YAO Meng[①②] JIA Kebin[①③④] SIU Wanchi[②]

[①](*Faculty of Information Technology, Beijing University of Technology, Beijing* 100124, *China*)

[②](*Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China*)

[③](*Beijing Laboratory of Advanced Information Networks, Beijing* 100124, *China*)

[④](*Advanced Innovation Center for Future Internet Technology, Beijing* 100124, *China*)

Abstract: The visual-based scene recognition and localization module is widely used in vehicle safety system. This paper proposes a new method of scene recognition based on local key region and key frame, which is based on the problem of large amount of training data, large matching complexity and low tracking precision. The proposed method meets the real-time requirements with high accuracy. First, the method uses the unsupervised method to extract the significant regions of the single reference sequence captured by the monocular camera as the key regions. The binary features with low correlation in key regions are also extracted to improve the scene matching accuracy and reduce the computational complexity of feature generation and matching. Secondly, key frames in the reference sequence are extracted based on the discrimination score to reduce the retrieval range of the tracking module and improve the efficiency. Practical field tests are done on real data of the light railway system in Hong Kong and the open test data set in Nordland. The experimental results show that the proposed method achieves fast matching and the precision is 9.8% higher than SeqSLAM which is based on global feature.

Key words: Visual-based localization; Key region; Key frame; Binary feature

收稿日期：2017-10-31；改回日期：2018-05-21；网络出版：2018-07-12

*通信作者： 贾克斌 kebinj@bjut.edu.cn

基金项目：国家自然科学基金面上项目(61672064)，北京市自然科学基金重点项目(KZ201610005007)

Foundation Items: The National Natural Science Foundation of China (61672064), The Beijing Natural Science Foundation (KZ201610005007)

1 引言

车辆定位系统为驾驶辅助系统以及车辆调度系统提供车辆的位置信息。车辆位置信息的准确度和实时性直接影响车辆行驶安全以及车辆运行效率。目前，全球定位系统(Global Positioning System, GPS)被广泛用于汽车与列车定位系统中，其精度能够达到约5 m，适用于大尺度定位系统，如普通城际列车的调度系统。与普通列车不同，城市轻轨列车的运行环境往往较为复杂，包括野外、街道、楼宇之间甚至室内，基于GPS的列车定位系统接收的信号容易受到建筑物影响而导致定位漂移，无法提供准确的车辆位置信息[1]。

近年来，视觉信息在定位系统中起着重要的作用，被广泛应用于车辆和移动机器人导航系统[1–3]。基于视觉信息的轻轨定位系统通过将当前帧的视频信息与数据集中的参考视频信息进行匹配，以获得列车的位置信息。然而在实际情况中，图像匹配往往受到光照变化等噪声干扰。为应对场景剧烈变化带来的干扰、提取场景稳定的特征，红外传感器被用来获取不受光照影响的长波热能图像[4]，激光雷达[5]以及双目摄像机[6]被用来获取场景场景的3维结构信息与深度信息。这些方法取得了较好的定位结果，但是其依赖特殊的传感器，因而算法移植性较差。因此如何在视频帧中提取稳定的视觉特征信息成为研究热点[1]。

在传统匹配算法中，特征点检测与描述算法被广泛用于稳定的图像特征提取，如旋转与尺度不变特征变换(Scale-Invariant Feature Transform, SIFT)[7]、加速稳健特征(Speed-Up Robust Feature, SURF)[8]、加速分割检测特征(Features from Accelerated Segment Test, FAST)[9]、二进制鲁棒独立基本特征(Binary Robust Independent Elementary Features, BRIEF)[10]及其改进算法ORB(Oriented FAST and Rotated BRIEF)[11]。这些算法主要利用场景的局部信息，被广泛应用于匹配定位系统，但无法在光照、视角发生变化的场景中取得较高的匹配准确度[12]。

近年来，一些针对视觉定位系统的特征提取方法被陆续提出。场景签名[12]利用丰富的数据集训练特征检测器，这些数据包含在不同条件下获取的同一场景，如晴天、雨天、下雪以及深夜。Han等人[13]提出了一种共享式外形表示学习方法(Shared Representative Appearance Learning, SRAL)，该方法融合了多种图像特征，并设计实现了一种基于视觉信息的移动车辆定位算法。Carlevaris-Bianco等人[14]使用300万个训练样本跟踪图像中不随时间变化的稳定特征。卷积神经网络(Convolutional Neural Network, CNN)[15]被用于训练一个地标检测器来识别稳定特征，以提高其视点不变性和条件不变性[16]。在这类算法中，文献[17]提供了类似对象的区域作为地标的候选者，并由CNN生成的特征从这些候选者中提取潜在的地标。Arroyo等人[18]基于卷积神经网络技术，设计提出了一种视觉拓扑定位算法(Convolutional Neural Network for Visual Topological Localization, CNN-VTL)，该方法使用大量数据训练得到场景特征，并基于该特征实现了一种针对车辆的视觉定位系统。这类基于样本学习方法需要大量采集场景信息并进行人工标定，因此基于单一参考序列的车辆定位算法仍然面临诸多挑战[19]。

由于列车往往在野外运行，其运行场景在一段时间内总是相似的，直接使用全局图像特征进行场景匹配并不能准确地估计当前位置。本文提出一种基于关键区域的二值化特征提取方法，并搭配单目摄像机实现了一套针对轻轨的实时定位系统。本文在下述创新点的基础之上实现了一套针对单目摄像机的单参考序列的轻轨实时定位系统：(1)提出了一种适用于场景匹配跟踪的关键区域与关键帧提取方法；(2)采用一种独立无监督的方法提取关键区域的二进制特征，极大减少了场景匹配的计算复杂度。该系统既不需要庞大的训练数据集，也不依赖任何特殊的传感器，便于移植到手机等具有视频采集以及计算能力的移动设备上。

本文的结构如下：第2节详细介绍了本文提出的方法，包括关键信息提取与二值化特征提取两部分。第3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。

2 基于视觉信息的轻轨定位系统

系统分为离线与在线两部分，系统结构如图1所示。针对参考帧相似度高的问题，离线模块为每个参考帧提取显著性高的区域作为关键区域，并以此提取参考视频中的关键帧。当车辆在与参考序列相同的路径上运行时，在线模块在关键帧提供的检索范围内，基于关键区域的二值化特征实时匹配当前帧与参考帧，由此可以提供列车的实时位置。

2.1 关键区域与关键帧的提取

2.1.1 关键区域提取 用于场景匹配的区域应包含较为稳定的视觉信息，这些信息不应随时间的变化而改变。因此，需要建立感兴趣区域(Region Of Interest, ROI)以减小非感兴趣区域中不稳定信息的影响。如图2所示，本文方法将列车采集视频中的非感兴趣区域进行移除，主要包括3类不稳定区

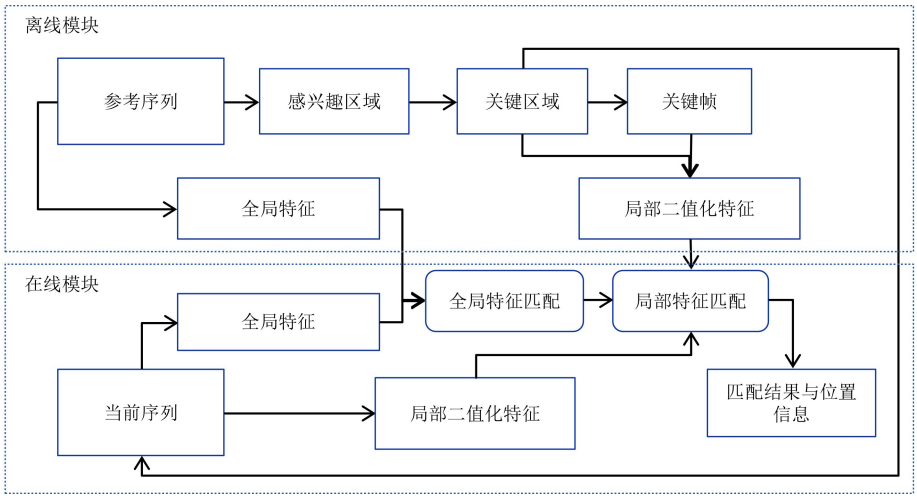
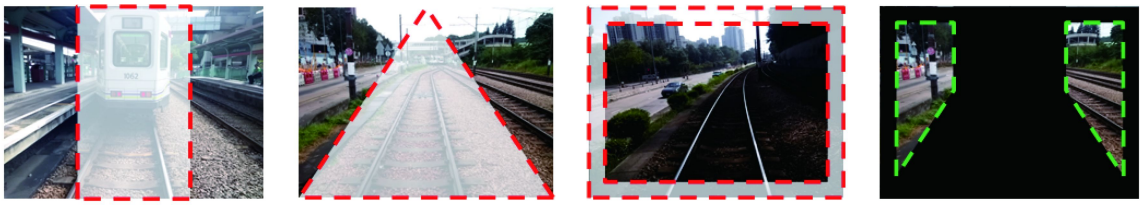


图 1 场景匹配系统流程图



(a) 运动物体区域　(b) 铁轨区域　(c) 边缘模糊区域　(d) 感兴趣区域

图 2 不稳定区域与感兴趣区域

域：运动物体、铁轨和视频帧边缘的动态模糊区域。首先，为避免与前车车距较近时前车对画面造成较大面积遮挡带来的干扰，去除视频帧中轴线两侧各150像素的区域。其次，去除铁轨信息，定义为视频顶边中心点与底边形成的三角形区域。最后，去除视频帧边缘40像素宽的模糊、扭曲区域。

感兴趣区域中所有像素记录了场景的整体信息，这种宏观上的信息可以定位车辆的大概位置，例如停靠站台或行驶路段。只有特定区域包含的显著性信息有助于为车辆提供更精确的位置信息，称为关键区域。显著性分数被用来衡量帧内区域的显著性程度，分数越高表示该区域越显著，反之则越普通。本文方法使用滑动窗口在ROI中对视频帧进行采样，计算ROI中的区域显著性分数。

记当前待计算显著性分数的视频帧为f_t，其时域邻域内包含F个视频帧，以图3为例，当前帧的邻域包含4个参考视频帧。当滑动窗口位于$(x,\ y)$位置时，如矩形框所示，分别计算窗口所包含的图像块$R(x,\ y,\ f_t)$与其他视频帧相同位置内图像块$R(x,\ y,\ f_{t'})$之间的差别，求和即得到当前帧该位置$(x,\ y)$的显著性分数：

$$S_{R(x,y,f_t)}=\frac{1}{N}\sum_{f_{t'}\in F,t'\neq t}D(R(x,y,f_t),R(x,y,f_{t'})) \quad (1)$$

其中，$D(R_A,\ R_B)$是图像块R_A和R_B之间的差别。$R(x,\ y,\ f_{t'})$是其他序列帧$f_{t'}$中相同位置的图像块。N是用来对比的视频帧数量，图3中该值为4。$S_{R(x,y,f_t)}$

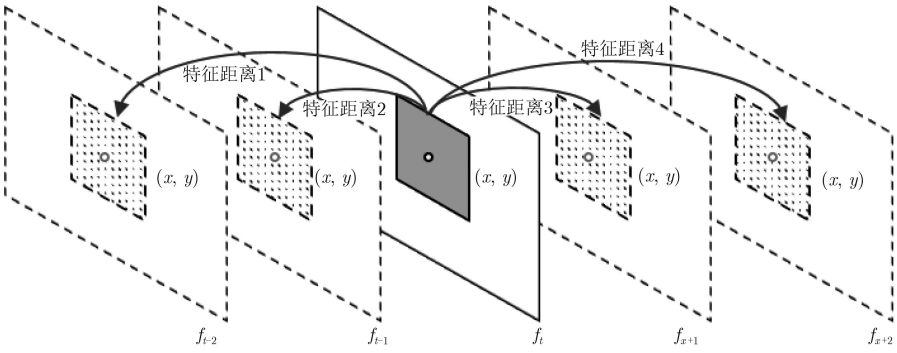
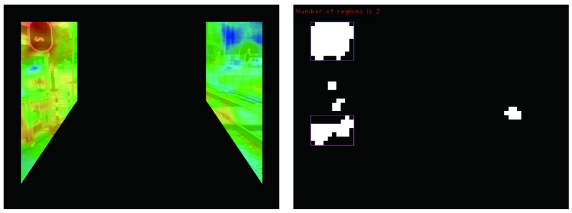


图 3 通过欧氏距离求和计算像素的显著性分数

是最终求和所得到的显著性分数。

图像块之间的差别可以用差帧图像的像素和或者图像特征之间的欧式距离来衡量。本算法使用梯度方向直方图(Histogram of Oriented Gradients, HOG)[20]特征计算图像块区别，以减少光线带来的影响。

显著性分数揭示了图像块的显著程度。如图4(a)所示，感兴趣区域内区域的显著性分数由不同颜色绘制。红色代表该区域具有较高显著性分数，蓝色反之。显著性分数高于一定阈值T_k的区域将被视为关键性区域，如图4(b)白色区域。



(a) 显著性分数 (b) 关键性区域

图 4 像素显著性分数与关键性区域

2.1.2 关键帧提取 某一帧的显著性分数定义为帧内所有区域的显著性分数总和。关键帧是一段时间内视频序列中包含特征信息最多的视频帧，其显著性分数应高于其他普通帧，且在全局尺度和某个时间邻域内都应具有较高显著性分数，以便匹配模块能够基于关键区域获得高可信度的匹配。关键帧不仅仅在全局尺度上具有较高显著性分数，并且在某个时间邻域内也具有较高的显著性分数。本文方法中的关键帧提取包含两步骤，首先提取具有局部最大显著性分数的视频帧，并将这些帧按照显著性分数降序排序，之后取前N_k帧作为关键帧。

2.2 无监督学习的二值化特征提取

基于本文提出的关键区域进行匹配大幅度减少了场景匹配计算量，降低了计算复杂度。在此基础上利用局部高分辨率视觉信息可获得更精确的匹配结果。局部视觉信息可以用传统基于统计学的方法提取，如HOG或SIFT特征。然而，这些方法计算复杂度高，难以在实时系统中使用。同时，该类浮点型特征描述符使用欧式距离计算特征相似度，导致匹配过程非常费时。为提高特征提取以及匹配过程的效率，多种二值化特征描述符被提出，如BRIEF和ORB，该类方法主要用于矩形内的一般局部图像块特征描述。其中，ORB特征基于角点检测，使用学习的方法得到二值化特征抽取模式。本文中提出了一种基于学习的二进制特征描述符，对不规则关键区域具有较高的描述力。该方法包括一种新颖的显著程度分析与贪婪算法。

二进制描述符可以通过级联一系列的像素对的二值化对比结果来获得。描述力强的二进制特征意味着具有可以增加当前帧与其相邻帧之间的差别，本文使用一种基于学习的方法提取描述力最强的像素对。显著性分数是用来评估像素对P的辨别度，P的显著性分数计算公式如式(2)：

$$S(P,F_q)=\sum_{i=0}^{M}(D(P,F_t)-D(P,F_q)) \tag{2}$$

其中，$S(P,\ F_q)$是当前帧内F_q某点对P的显著性分数，$D(P,\ F_q)$是当前查询帧F_q内点对P的两个像素之间的灰度差，$D(P,\ F_i)$是第i个相邻帧内点对P的两个像素之间的灰度差。M是相邻帧的数量。

将所有像素对按照显著性分数降序排序，前N个像素对可以认为是当前帧中显著性最高的点对选择模式。然而，实验表明，使用该方法筛选出的点对虽然显著性分数都比较高，但是往往点对都集中出现在近似位置，从而使得二值化结果具有较高的相关性。例如，如果选择显著性分数高的像素对$P((x_1,\ y_1),\ (x_2,\ y_2))$，其空间相邻位置的点对$P((x_1{+}1,\ y_1{+}1),\ (x_2{+}1,\ y_2{+}1))$可能有类似的高分数进而被选择，导致基于显著性分数的像素对生成的二进制描述符的信息量减少。因此，应进一步筛选相关性低的点对。

主成分分析(Principal Component Analysis, PCA)被广泛应用于分析提取数据的主要维度。由于本系统只包含少量训练样本，不能提供主成分分析提取所需足够的维度数，因而无法使用主成分分析的方法筛选相关性低的点对。贪婪算法可以检查关键区域内所有可能的像素对，进而从高显著性分数的点对中进一步筛选出相关性低的像素对。

本文提出的基于互相关系数的贪婪算法可以提取优质的二值化特征。贪婪算法的训练样本集合由当前帧与其邻域内的若干帧组成。首先计算训练样本集合中所有可能的像素对的灰度差，建立训练矩阵$\boldsymbol{T}$。假设当前帧包含关键区域K个像素和M个相邻帧，因此点对共有$K\times(K{-}1)/2$种模式，训练矩阵包含$K\times(K{-}1)/2$行和$M{+}1$列。训练矩阵中的每一行代表一个像素对训练集的像素灰度差的分布。迭代训练过程如下：

(1)将训练矩阵$\boldsymbol{T}$内所有行按显著性分数降序排列，因此显著性分数最高的像素对位于第1行；

(2)将$\boldsymbol{T}$中的第1行被移动到结果矩阵$\boldsymbol{R}$中初始化结果矩阵，并将该第1行从$\boldsymbol{T}$中删除；

(3)提取$\boldsymbol{T}$的第1行，并计算这一行和$\boldsymbol{R}$中的所有行的相关性；

(4)若所有相关小于预定义的阈值C，则将该行放入$\boldsymbol{R}$，并从$\boldsymbol{T}$中删除此行，转到第(3)步，否则，直接从$\boldsymbol{T}$中删除此行并转到第(3)步；

(5)若$\boldsymbol{R}$中的行数达到预定义的N，停止迭代；

(6)若$\boldsymbol{T}$为空，停止迭代。

3　实验结果与分析

在实验中，使用本文提出的方法提取的关键区域、关键帧和基于机器学习的二进制场景特征描述方法被用于匹配测试序列与参考序列，实验使用台式计算机作为处理平台。

3.1　数据集

实验使用了来自香港港铁(Mass Transit Railway, MTR)提供的轻轨数据集以及挪威广播公司(Norwegian Broadcasting Corporation, NRK)公开的Nordland数据集[21]。

香港轻轨数据集采集自轻轨507号路线，共包含3组视频序列，视频分辨率640×480像素，帧率为25帧/s，共包含13859帧。这些序列由安装在轻轨车辆上的单目相机拍摄。每组视频序列包含2段序列，这2段序列采集自同一列火车在不同时间运行在相同的路径上。Nordland数据集包含4个序列分别在春夏秋冬4个季节中拍摄，其分辨率为1920×1080像素，帧率为25帧/s，本文从中选取10000帧作为训练和测试数据。4个序列由人工逐帧匹配，即这4个序列中帧号相同的帧拍摄于相同的位置。

3.2　单帧场景识别评估

在参考序列中提取的关键区域和关键帧被用于场景识别与跟踪锁定。所以，关键帧与关键性区域的有效性将由场景识别的质量来衡量。其中，当列车再次捕捉到与关键帧内容相同的视频内容时，当前帧与关键帧可以提供高可信度匹配分数以确定列车位于关键帧位置。

实验比较了以下4种不同的场景描述方法，包括基于全局特征的方法和基于局部关键区域的方法。在匹配过程中，统一使用HOG特征，评价标准为匹配结果和真实标定位置(Ground truth)的平均错误偏移。

全局特征即将视频帧当做一个整体使用一个HOG特征作为描述符，计算图像之间的差别。在第2组中，为保留画面的内容的相对位置信息，将视频帧分为40×40的互不重叠的宏块，分别计算每个宏块的HOG特征。匹配时，计算当前帧与参考帧每个宏块之间的差别，并对所有这些差求和得到全帧的差别。为评估本文中的感兴趣区域的性能，在第3组中，匹配时只考虑感兴趣区域内的宏块。最后一组为本文提出的关键区域，该方法计算每个连通的关键区域的HOG特征。

匹配当前序列与参考序列后，对于每个当前帧，最匹配的参考帧序号和实际标定参考帧序号之间的差被称为错误偏移，其绝对值的平均值被用来评价4种方法的匹配精度，该平均错误偏移的单位为帧。理想情况下此偏移量接近零，意味着所有匹配结果与真实标记结果完全一致。如表1所示，本文提出的基于关键区域方法的平均错误偏移最低。基于宏块HOG特征方法的时间复杂度最高，其每帧图像匹配时间达到62.42 s。使用感兴趣区域减少了场景匹配计算时间，但平均错误偏移上升了0.16帧。与全局HOG特征方法相比，本文提出的基于关键区域的HOG特征在计算时间代价与匹配质量之间提供了一种折中的方案。

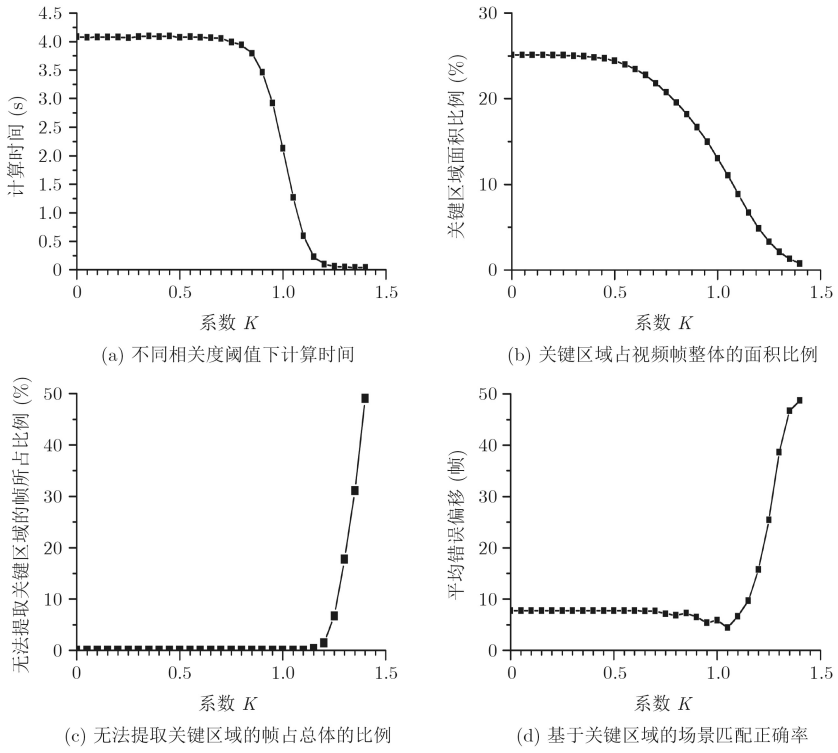
表 1　时间复杂度与算法平均错误偏移

对比方法	平均错误偏移(帧)	时间(s)
全局HOG特征	15.24	0.0593
基于宏块HOG特征	2.10	62.4205
基于感兴趣区域内宏块HOG特征	2.26	13.5960
本文基于关键区域HOG特征	1.44	3.6058

如2.1节所述，当前帧的关键区域由显著性分数和预定义的阈值T_k决定。在本文算法中，阈值T_k采用自适应的阈值，因为每帧中的显著性分数分布在不同尺度，因而无法使用统一的绝对阈值。例如，关键帧中的区域的显著性分数远高于非关键帧。因此，使用系数K间接调整阈值T_k，T_k是帧内平均显著性分数与系数K的乘积。

首先测试系数K不同取值下场景识别的时间代价，系数K取值范围从0到1.40。如图5(a)中所示，随着系数K从0.75升高，场景识别的计算时间代价迅速下降。因此，较大的系数K的值能够将该匹配方法应用于实时系统中的单个场景识别。其主要原因是当系数K取较大值时，只有少量区域被定义为关键区域。本系统使用的HOG特征其基本单元格的大小固定为10像素，使用较小的关键区域意味着描述区域中单元格数量减少从而降低HOG特征计算复杂度。

图5(b)中展示了不同系数K的取值下，关键区域面积占整个视频帧的百分比。该组实验记录了每帧关键区域面积与整个视频帧面积的比例。纵轴是数据集中所有帧的关键区域的平均比例。可以看出，关键区域的百分比随K系数上升而下降。当



(a) 不同相关度阈值下计算时间　(b) 关键区域占视频帧整体的面积比例

(c) 无法提取关键区域的帧占总体的比例　(d) 基于关键区域的场景匹配正确率

图 5 不同系数K对关键区域的影响

K=0时，意味着感兴趣区域内所有像素都被视为重点区域。在这种情况下，算法使得匹配系统专注于帧内25%的区域，而不是整个视频帧。

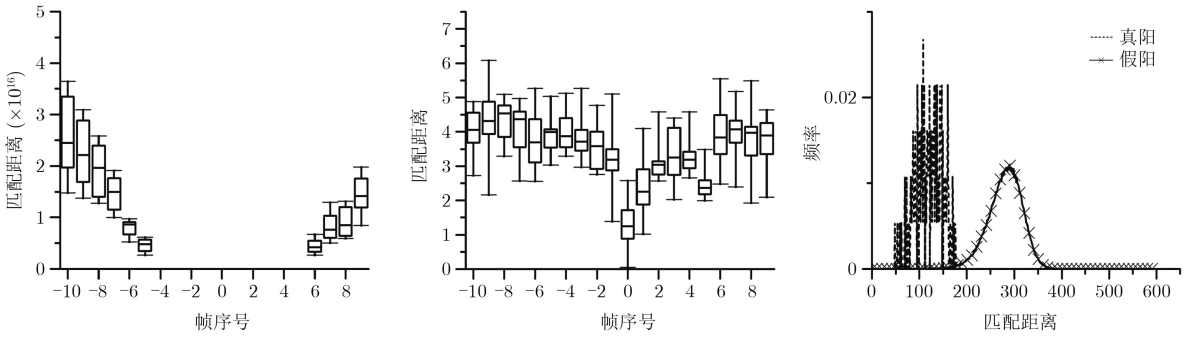
计算时间代价的趋势和关键区域面积百分比揭示了系数K取值较大时其运算效率将随之提高。当系数K从0增加到1.40时，场景识别的计算时间从4 s下降到0.039 s，同时，关键区域的百分比从25%降至0.76%。然而，当系数K取值过大时，增加了部分帧无法找到关键区域的风险。图5(c)所示，当系数K的取值大于1.10时，数据集内的一些帧将无法检测出关键地区。使用较大的K取值，导致没有关键区域的帧的百分比增加，这些没有任何关键区域的帧需要使用感兴趣区域内所有区域参与特征计算，降低了场景识别的效率。

为了确定最合适的系数K的取值，该部分使用了所有可能的系数K的取值在数据集中进行场景匹配计算并统计平均错误偏移。同时，为保证结果的可信度，实验对比了不同关键区域的实验结果，7组关键区域通过更改式(1)中的参数N获得，N的变化范围为5～50。图5(d)显示每种参数选择下场景匹配的错误率，纵轴以平均错误偏移来表示该错误率。由图可知，当系数K取值过大，则匹配错误随之急剧升高，如K=1.40。该结果表明，虽然系统选择了最显著的关键区域，但是由于所选的关键区域包含的像素过少导致视觉信息受限，从而使得场景匹配容易受到噪声干扰。另一方面，当系数K=1.05时，其错误率最低。这一结果为系数K的取值提供了参考，该取值既保证了场景识别的可靠性，又保证了算法的运行效率。3.3节使用系数K=1.05作为阈值提取关键区域，并进行基于机器学习的二值化特征的提取过程。

3.3 多帧场景跟踪评估

多帧场景跟踪使用Nordland数据集进行测试。跟踪模块首先使用SeqSLAM[21]算法匹配当前序列与参考序列，获得与当前帧匹配上的候选参考帧集合。SeqSLAM作为一种场景序列匹配方法被广泛应用于基于路径的视觉定位算法中[22–24]。SeqSLAM中的全局特征由降采样的归一化图像生成，其分辨率为32×24像素。帧间相似度采用当前帧与参考帧之间的第1范数距离作为衡量标准，两帧之间的距离低意味着它们具有相似的外观。在候选参考帧集合中，跟踪模块使用二值化特征进行2次验证，验证结果以匹配分数为依据。匹配分数越高，意味着两帧于同一个地方采集的可能性越高。准确率用于评估跟踪模块的性能。真阳性样本(True Positive, TP)被定义为真实标定位置附近3帧范围内的阳性样本；否则，阳性样本被认为是假阳性样本(False Positive, FP)。准确率可以通过TP/(TP+FP)计算得到。

图6(a)显示SeqSLAM中13个错误匹配帧的匹配距离分布，图中记录了真实标定位置周围20帧的距离分布。纵轴是匹配距离，横轴是邻近帧与真实



(a) SeqSLAM 的匹配距离的分布　(b) 基于二值化特征的匹配距离的分布　(c) 基于二值化特征的匹配距离频率分布

图 6 本文方法与SeqSLAM在高帧率场景匹配中的结果对比

标定位置的相对索引。由图可知，在SeqSLAM的匹配结果中，距离真实标定位置前后约10帧范围内的参考帧拥有相似匹配距离，这使得SeqSLAM跟踪模块难以确定与当前帧最佳匹配的参考帧。同时，使用本文提出的二值化特征验证这些帧时，匹配分数的峰值总是出现在真实标定位置，如图6(b)所示。这表明基于机器学习的二值化特征描述方法只在真实标定位置匹配距离较低，使得匹配模块能够提供更准确更可信的跟踪结果。

此外，为进一步验证本文所提出的二值化特征的性能，实验统计了对于SeqSLAM方法的真阳性样本和假阳性样本其二值化特征匹配方法匹配分数的分布情况，如图6(c)所示。图中虚线是真阳性样本不同匹配分数的出现频率，带叉实线对应假阳性样本的频率。这两条曲线在匹配距离为75处明显分隔开来。因此，本文提出的方法可以区分高度相似的视频帧，提高SeqSLAM跟踪器的跟踪准确率。

表2统计了两种跟踪算法的准确率、匹配偏移及匹配时间。经过本文提出的二值化特征校验，跟踪准确度被提高至99.36%。匹配偏移量下降了36.07%，同时并未显著增加场景匹配时间。该结果表明本文提出的基于机器学习的二值化特征能够提供场景中更显著的视觉信息，从而得到更加精确的匹配，而SeqSLAM中的全局特征只能提供较为粗略的匹配结果。

4　结论

基于单目摄像机的轻轨实时定位系统具有采集设备简单，推广应用前景好的优势。随之而来的主要困难是系统计算复杂度高、准确度低的问题。对此，本文提出了一种基于关键区域与无监督学习的轻轨定位方法。该方法所提取的关键区域不仅提高了单个场景识别的精度，同时降低了计算时间成本。其次，本文设计了一种新的显著性衡量标准，实现了在关键区域内的二值化图像特征提取。最终，本文实现了一套基于单视点视频的轻轨实时定位系统。实验结果表明，基于机器学习的二值化特征方法不仅提高了场景匹配精度，其极低的计算时间代价使其得以在基于场景跟踪的车辆实时定位系统中应用。

表 2　场景跟踪准确率、匹配偏移及匹配时间

	SeqSLAM	本文基于二值化特征的场景跟踪方法	△(%)
准确率(%)	89.56	99.36	+9.80
匹配偏移(帧)	1.3652	0.8728	–36.07
匹配时间(ms)	53.23	54.82	+2.99

参 考 文 献

[1] LOWRY S, SUNDERHAUF N, NEWMAN P, *et al*. Visual place recognition: a survey[J]. *IEEE Transactions on Robotics*, 2016, 32(1): 1–19. doi: 10.1109/TRO.2015.2496823.

[2] DAYOUB F, MORRIS T, BEN U, *et al*. Vision-only autonomous navigation using topometric maps[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2013: 1923–1929. doi: 10.1109/IROS.2013.6696611.

[3] MURILLO A C, SINGH G, KOSECKÁ J, *et al*. Localization in urban environments using a panoramic gist descriptor[J]. *IEEE Transactions on Robotics*, 2013, 29(1): 146–160. doi: 10.1109/TRO.2012.2220211.

[4] MADDERN W and VIDAS S. Towards robust night and day place recognition using visible and thermal imaging[OL]. https://eprints.qut.edu.au/52646/.

[5] MCMANUS C, FURGALE P, and BARFOOT T D. Towards lighting-invariant visual navigation: An appearance-based approach using scanning laser-rangefinders[J]. *Robotics and Autonomous Systems*, 2013, 61(8): 836–852. doi: 10.1016/j.robot.2013.04.008.

[6] LINEGAR C, CHURCHILL W, and NEWMAN P. Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera[C]. 2016 IEEE International Conference on. Robotics and Automation (ICRA),

Stockholm, Sweden, 2016:787–794. doi: 10.1109/ICRA.2016.7487208.

[7] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91–110. doi: 10.1023/B:VISI.0000029664.99615.94.

[8] BAY H, ESS A, TUYTELAARS T, *et al*. Speeded-up robust features (SURF)[J]. *Computer Vision and Image Understanding*, 2008, 110(3): 346–359. doi: 10.1007/11744023_32.

[9] ROSTEN E, PORTER R, and DRUMMOND T. Faster and better: A machine learning approach to corner detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(1): 105–119. doi: 10.1109/TPAMI.2008.275.

[10] CALONDER M, LEPETIT V, STRECHA C, *et al*. BRIEF: Binary robust independent elementary features[C]. Proceedings of the 11th European Conference on Computer Vision, Crete, Greece, 2010: 778–792. doi: 10.1007/978-3-642-15561-1_56.

[11] RUBLEE E, RABAUD V, KONOLIGE K, *et al*. ORB: An efficient alternative to SIFT or SURF[C]. Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 2011: 2564–2571. doi: 10.1109/ICCV.2011.6126544.

[12] MCMANUS C, UPCROFT B, and NEWMANN P. Scene signatures: Localised and point-less features for localisation[C]. Robotics:Science and Systems, Berkeley, USA, 2014: 1–9. doi: 10.15607/rss.2014.x.023.

[13] HAN Fei, YANG Xue, DENG Yiming, *et al*. SRAL: Shared representative appearance learning for long-term visual place recognition[J]. *IEEE Robotics and Automation Letters*, 2017, 2(2): 1172–1179. doi: 10.1109/LRA.2017.2662061.

[14] CARLEVARIS-BIANCO N and EUSTICE R M. Learning visual feature descriptors for dynamic lighting conditions[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, USA, 2014: 2769–2776. doi: 10.1109/IROS.2014.6942941.

[15] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012: 1097–1105. doi: 10.1145/3065386.

[16] SÜNDERHAUF N, SHIRAZI S, JACOBSON A, *et al*. Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free[C]. Proceedings of Robotics: Science and Systems XII, Rome, Italy, 2015: 296–296. doi: 10.15607/rss.2015.xi.022.

[17] ZITNICK C L and DOLLÁR P. Edge boxes: Locating object proposals from edges[C]. Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 391–405. doi: 10.1007/978-3-319-10602-1_26.

[18] ARROYO R, ALCANTARILLA P F, BERGASA L M, *et al*. Fusion and binarization of CNN features for robust topological localization across seasons[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Daejeon, South Korea, 2016: 4656–4663. doi: 10.1109/IROS.2016.7759685.

[19] LINEGAR C, CHURCHILL W, and NEWMAN P. Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera[C]. IEEE International Conference on Robotics and Automation, Stockholm, Sweden, 2016: 787–794. doi: 10.1109/ICRA.2016.7487208.

[20] DALAL N and TRIGGS B. Histograms of oriented gradients for human detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005: 886–893. doi: 10.1109/CVPR.2005.177.

[21] MILFORD M J and WYETH G F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights[C]. IEEE International Conference on Robotics and Automation, Saint Paul, USA, 2012: 1643–1649. doi: 10.1109/ICRA.2012.6224623.

[22] BRESSON G, ALSAYED Z, LI Yu, *et al*. Simultaneous localization and mapping: A survey of current trends in autonomous driving[J]. *IEEE Transactions on Intelligent Vehicles*, 2017, 2(3): 194–220. doi: 10.1109/TIV.2017.2749181.

[23] KIM P, COLTIN B, ALEXANDROV O, *et al*. Robust visual localization in changing lighting conditions[C]. IEEE International Conference on Robotics and Automation, Singapore, 2017: 5447–5452. doi: 10.1109/ICRA.2017.7989640.

[24] BAI Dongdong, WANG Chaoqun, ZHANG Bo, *et al*. Sequence searching with CNN features for robust and fast visual place recognition[J]. *Computers and Graphics*, 2018, 70: 270–280. doi: 10.1016/j.cag.2017.07.019.

姚 萌：男，1988年生，博士生，研究方向为图像处理、视觉信息定位.

贾克斌：男，1962年生，教授，研究方向为图像/视频信号与信息处理技术、生物信息处理与计算技术、基于Internet网的多媒体系统等.

萧允治：男，1950年生，教授，研究方向为图像处理、小波变换、模式识别等.