DOI: 10.13873/J.1000-9787(2019)05-0125-04

动态环境下结合语义的鲁棒视觉 SLAM*

王金戈[1,2]，邹旭东[1,2]，仇晓松[1,2]，蔡浩原[1]
(1.中国科学院 电子学研究所 传感技术国家重点实验室，北京 100190；
2.中国科学院大学 电子电气与通信工程学院，北京 100049)

摘　要：针对传统同时定位与地图构建(SLAM)在动态环境中受动态物体干扰而导致精度低、鲁棒性差的缺点，提出了一种结合语义的鲁棒视觉 SLAM 算法。采用深度学习技术构建基于卷积神经网络的物体检测器，结合先验知识，在语义层面实现对动态物体的检测；提出基于速度不变性的相邻帧漏检补偿模型，进一步提高物体检测网络的检出率；构建基于特征点的视觉 SLAM 系统，在跟踪线程中对动态物体特征点进行剔除，以减小错误匹配造成的位姿估计的误差。经实验验证：系统在极端动态环境测试中保持定位不丢失，在 TUM 动态环境数据集测试中，定位精度比 ORB-SLAM2 提高 22.6%，性能提高 10%。

关键词：视觉同时定位与地图构建(SLAM)；动态环境；语义；物体检测
中图分类号：TP24　　文献标识码：A　　文章编号：1000-9787(2019)05-0125-04

Robust visual SLAM with semantics in dynamic environment*

WANG Jinge[1,2], ZOU Xudong[1,2], QIU Xiaosong[1,2], CAI Haoyuan[1]
(1. State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China; 2. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Aiming at the shortcomings of low precision and poor robustness of traditional simultaneous localization and mapping(SLAM) in dynamic environment due to the disturbance of dynamic objects, a robust visual SLAM algorithm based on semantic is proposed. Deep learning is used to construct object detector based on convolutional neural network(CNN), combined with prior knowledge, so as to realize the detection of dynamic object on the semantic level. Propose a compensation model for missed detection in adjacent frames based on the invariance of speed, which further improve the detection rate of object detecting network. A feature point-based visual SLAM system is constructed, in tracking line path, feature points on dynamic objects are removed so that the error of pose estimation is reduced. The experiment verifies that in extreme dynamic environment test, the system keeps positioning without losing. In TUM dynamic environment data set test, compared with ORB-SLAM2, the positioning precision is improved by 22.6%, and the performance is improved by 10%.

Keywords: visual simultaneous localization and mapping(SLAM); dynamic environment; semantic; object detection

0　引　言

传统的同时定位与地图构建(simultaneous localization and mapping, SLAM)技术建立在静态环境下，不考虑环境物体的运动。而实际环境中，人的走动、车辆的来往都会造成环境动态变化，从而使 SLAM 系统建立的地图无法保持长时间的一致性，基于视觉的特征也会因为物体的运动而变得不稳定。在仓储、无人驾驶等定位精度要求较高的领域，移动中的人对 SLAM 定位精度的影响会导致位姿漂移、跟踪失败、误差累积等问题。

为了使 SLAM 在动态环境下正常工作，需要避免使用处于动态物体上的特征点，因此，需要事先计算出动态物体的位置。目前常用的动态物体提取方法都是基于几何特征[1,3]，当面对更加极端的动态环境时，如人靠近镜头的走动，依然会失效。

本文提出了一种结合语义的鲁棒视觉 SLAM——Dynamic-SLAM 算法，采用深度学习技术在语义层面实现对动态物体的检测，并对动态物体特征点进行剔除，以消除其在 SLAM 定位与建图中的误差。实验证明，该方案在各种动态环境中都取得了较好的定位精度和鲁棒性。

1　系统框架

本文在ORB-SLAM2[3-5]的基础上，增加基于语义的动

收稿日期：2018-01-23
* 基金项目：国家青年千人基金资助项目；国家自然科学基金资助项目(61372052)

态物体判定模型，并优化基于特征点的视觉里程计算法，使其能够舍弃附着在动态物体上的特征点，只采用非动态物体的特征点参与位姿估计和非线性优化，从而避免动态物体特征点的干扰。在物体检测部分，提出了基于运动模型的物体检测补偿算法，进一步提高了物体检测精度。

单目相机实时采集的图像作为 SLAM 定位与建图模块和物体检测模块的输入，物体检测模块的输出经过语义校正模块后实时反馈给 SLAM 定位与建图模块，SLAM 定位与建图模块最后给出定位和建图结果。

1.1 物体检测

本文采用文献[10]提出的 SSD(single shot multibox detector)物体检测网络，该网络使用 VGG16 的基础网络结构，保留前 5 层不变，利用 Atrous[11] 算法将 fc6 和 fc7 层转换成 2 个卷积层，再在后面增加 3 个卷积层和 1 个平均池化层。使用不同网络层的信息来模拟不同尺度下的图像特征，最后通过非最大抑制得到最终的检测结果。由于舍弃了最初的候选框生成阶段，使得整个物体检测流程能够在单一网络下完成，从而实现较高的检测效率(46 FPS，Titan X)和检测精度(77.2%)。

在动态环境 SLAM 中，动态物体检测的成功与否直接决定了系统的其他模块是否能够正常执行。一旦发生漏检，相邻两张图像间的巨大差异将会导致特征点数量急剧变化，从而导致系统的不稳定。为了能够稳定、有效地剔除动态特征点，必须在物体检测时获得足够高的检测精度。在常规的物体检测任务中，由于各个图片间不具有明显的关联，无法通过上下文信息提高检测精度。但在 SLAM 中，由于视频帧按照时间序列抵达，可以借助前若干帧的检测结果预测下一次的检测结果，从而弥补下一次可能出现的漏检或误检。基于这一思想，本文提出了相邻帧漏检补偿模型，该模型基于一个合理的假设“动态物体的运动速度不会超过某个阈值”。用 X 表示动态物体的坐标，V_{th} 表示动态物体运动速度的阈值，FPS 表示帧率，两者之间应该满足 $X < V_{\text{th}}/FPS$ 的关系。实践中需要将 V_{th} 设置为合适的值，太小会使系统过于敏感，导致正确检测被认为漏检，太大则可能使多个动态物体的检测区域重叠。

漏检补偿流程如下：

1) 当前帧 K_1 进入 SSD 网络，输出检测到的物体列表，列表中的每一项包括检测出的物体的位置坐标 $X_{1i}(0 < i < n_1, n_1$ 为 K_1 检测结果的数量)。

2) 若对于前一帧 K_0 的检测结果中的每一项 $X_{0j}(0 < j < n_0, n_0$ 为 K_0 检测结果的数量)，在当前帧检测结果中不存在 $|X_{1i} - X_{0j}| < V_{\text{th}}/FPS$，此时认为出现漏检，需要把 X_{0j} 添加进当前帧的检测结果列表中。

3) 修正后的检测结果列表作为动态物体判定原始数据。

1.2 动态物体判定

本文在语义的层面上提出了基于先验知识的动态物体判定方法。SLAM 系统如果不从语义层面理解周围的环境，就无法真正区分哪些是动态的，哪些是静态的，只能在短时间内找出运动的物体，而无法保证长时间的一致性。因此，本文将物体检测的结果与先验知识相结合，给出动态物体判定模型。根据人的先验知识，对物体的动态特性评分，0 分为静态物体，10 分为动态物体，常见物体在该区间上所处的大致位置如图 1 所示。

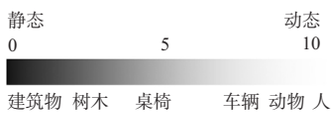


图 1 常见物体的动态特性评分

图中只列出了一部分常见的物体，其他物体的评分可以根据具体应用设定合适的分数。将物体分数与一个事先定义的阈值相比较，分数高于阈值时判定为动态物体，低于阈值时则判定为静态物体。阈值的大小视情况而定，通常可设为 5。

1.3 动态环境 SLAM

在 ORB-SLAM2 原有框架的基础上，增加了物体检测线程和基于语义的校正模块。新增的模块与 ORB-SLAM2 已有的 3 个线程的关系如图 2 所示。在新的视频帧抵达后，同时传入物体检测线程和跟踪线程，两者并行地对图像进行处理。物体检测线程采用前述的 SSD 物体检测网络计算出物体的类别和位置，进一步由基于语义的校正模块将其中的物体分为动态物体和静态物体，最后把动态物体的位置提供给跟踪线程。

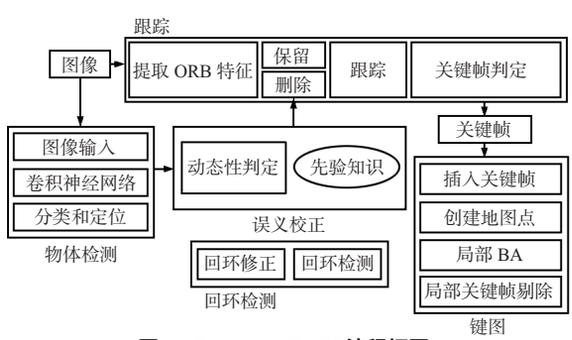


图 2 Dynamic-SLAM 流程框图

跟踪线程对每一帧图像提取 ORB 特征[12]，通过与参考帧的特征匹配，得到 2 张图像间特征点的对应关系，利用这些对应关系估计相机位姿。在初始化完成的情况下，相机位姿估计是一个 PnP(perspective-n-point)问题，求解该问题的方法有很多[13,14]，本文采用以集束调整(bundle adjustment)[15]为代表的非线性优化方法，该方法可以充分利用所有匹配结果，得到位姿的最优估计。构建非线性优化问题，最小化重投影误差如下

$$\xi^* = \arg\min \frac{1}{2} \sum_{i=1}^{n} \| u_i - \frac{1}{S_i} k \exp(\hat{\xi}) P_i \|_2^2$$

式中　ξ 将观测到的像素坐标与 3D 点按照当前位姿 ξ 投影后的 2D 坐标求差，该误差即为重投影误差。优化目标是找到一个相机位姿 ξ，使得重投影误差最小。

在动态环境中，受到运动物体的影响，动态物体上特征点的重投影误差会处于过高的水平，导致相机位姿 ξ 无法收敛到最优值，定位误差显著增大。为此 Dynamic-SLAM 的追踪线程在提取 ORB 特征后，根据当前检测到的动态物体的位置实施特征点剔除操作，将动态物体其上的特征点予以剔除。在后续的局部地图匹配、相机位姿估计和非线性优化中，只利用静态物体上的特征点，保证整个过程中重投影的一致性，使 SLAM 系统不受动态物体干扰。

为了保证 SLAM 系统的实时性，物体检测和跟踪分处 2 个线程，设计了安全高效且支持并发操作的数据结构 Detection来传递检测结果，并使用互斥锁 Unique_lock 保证不发生访问冲突，在写入操作执行前需事先获取锁。物体检测线程和跟踪线程的处理速度并不一致，采取异步读写共享变量的方式实现线程间通信，最大限度地利用 CPU 时间。

2　实验与结果分析

本文设计并实施了一系列实验来验证 Dynamic-SLAM 系统在动态环境下的鲁棒性和定位精度。使用 TUM RGB-D benchmark [16] 中的 Walking_rpy 数据集验证物体识别的准确率和检出率，使用 2 段采集的数据集验证动态物体干扰下初始化和定位的鲁棒性，使用 TUM RGB-D Benchmark 中的 Walking_xyz 数据集验证一般动态环境下的定位精度。

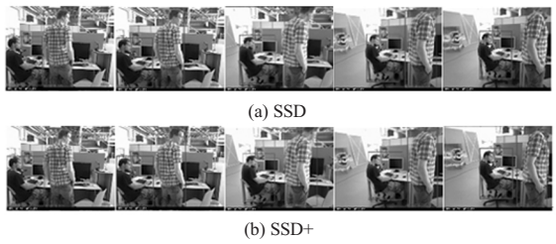
实验运行环境为 Intel Core i5-7300HQ(4 核 2.5 GHz) ，8 GB内存，NVIDIA GeForce GTX1050Ti 显卡，4 GB 显存。

2.1　物体检测

在该项测试中，采用 Walking_rpy 数据集，在连续487 帧图片中，SSD 的原始检测结果成功检出 401 次，失败 86 次，检出率 82.3 %。经过运动补偿后的检测结果成功检出 486 次，失败 1 次，检出率达到 99.8 %。图 3 所示为物体检测结果每隔 30 帧的抽样，左侧为 SSD 的原始检测结果，右侧为经过相邻帧漏检补偿后的检测结果，方框为检测到的物体位置，左下角标签标注了物体类别。实验表明，相邻帧漏检补偿模型大大提高了物体检测的检出率，为后续的 SLAM 定位与建图模块打下了良好的基础。

2.2　动态物体干扰下的初始化测试

初始化的成功与否关系到后续的定位是否准确。由于初始时刻尚无事先建立的地图，只能通过帧间匹配来确定相机的运动，导致动态环境下的初始化变得尤为困难。在动态物体的干扰下，SLAM 很容易错误初始化。实验中，将摄像头放置在桌面上固定不动，测试人的来往走动对初始化会造成怎样的影响。

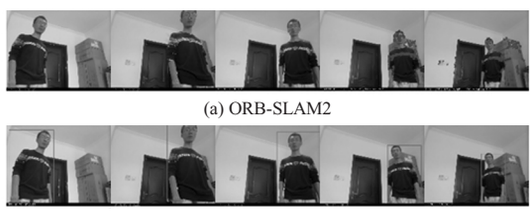


(a) SSD

(b) SSD+

图 3　物体检测测试

实验结果如图 4 所示。可以看到，ORB-SLAM 2 在面对动态物体时，不能分辨前景和背景物体，无法排除动态物体的干扰，特征点大多聚集在动态物体之上，到第 4 幅图时已经错误初始化。而 Dynamic-SLAM 成功检测出动态物体的位置，并将其上的特征点剔除，从而避免了错误初始化。



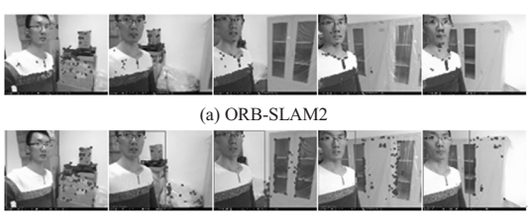
(a) ORB-SLAM2

(b) Dynamic-SLAM

图 4　动态物体干扰下的初始化测试

2.3　抗动态物体干扰测试

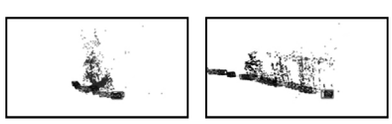
针对人始终存在于相机视野中的情况设计了一个测试集，分别测试 ORB-SLAM2 和 Dynamic-SLAM 的运行效果。如图 5 所示。



(a) ORB-SLAM2

(b) Dynamic-SLAM

图 5　抗动态环境干扰测试



(a) ORB-SLAM2　　(b) Dynamic-SLAM

图 6　抗动态环境干扰测试定位和建图结果

图 5(a) 中，为 ORB-SLAM 2 的特征提取(点状为特征点) ，图 5(b) 为 Dynamic-SLAM 的动态物体识别和特征提取。图 6(a) ORB-SLAM 2 的定位与建图结果，6(b) Dynamic-SLAM 的定位与建图结果。可以发现，由于无法区分动态物体和静态物体，ORB-SLAM 2 提取的特征点大多集中在人身上，使得系统把人当成背景环境，从而导致定位的结果完全依赖于人和相机之间的相对运动。在图 6(a) 中可以明显地看到，相机在运动一段时间后就完全停止了，地图点也不再更新。而 Dynamic-SLAM 能够自动选取静态环境

的特征，定位结果是一条连续的直线，建立的地图也基本与实际场景相吻合。

2.4 TUM 动态环境数据集测试

本文选择 TUM RGB-D benchmark 中的 Walking_xyz 数据集，该数据集的场景中有 2 个人在办公桌周围来回走动，运动幅度大，且在视野中占据了不小的比例，本文测试两种算法在该场景下的定位精度和性能。

图 7（a）为 ORB-SLAM2 的实时画面，图 7（b）为 Dynamic-SLAM 的实时画面。记录了所有关键帧的定位结果，与真实值比较并计算误差。

(a) ORB-SLAM2 (b) Dynamic-SLAM

图 7 Walking_xyz 数据集测试

从实验结果可以看出，在均方根误差、平均误差、误差中间值、误差标准差、最小误差和最大误差这 6 个指标中，Dynamic-SLAM（分别为 1.68，1.59，1.74，0.55，0.54，2.81 cm）都明显优于 ORB-SLAM2（分别为 2.17，2.05，2.02，0.68，0.95，4.01 cm）。以均方根误差为标准，Dynamic-SLAM 的精度比 ORB-SLAM 2 提高了 22.6 %。在性能方面，记录了 2 个算法的运行时间，多次运行取平均值的方法 Dynamic-SLAM 的性能比 ORB-SLAM 2 提高了10 %。虽然增加了物体检测流程，但由于物体检测放在一个独立线程中，并利用了 GPU 加速，非但没有拉低整个系统的运行速度，反而使运行时间缩短了。这得益于无效特征点的减少，使系统只借助于有效的特征点进行计算，从而节约了位姿估计和非线性优化的时间。

3 结 论

本文提出了一种动态环境下结合语义的鲁棒视觉 SLAM 算法，实验结果表明，改进后的算法在动态环境下的定位和建图精度更高，鲁棒性更强。与目前 State-Of-the-Art 的视觉 SLAM 算法 ORB-SLAM 2 相比，在动态环境数据集下的定位精度提高 22.6 %，性能提高 10 %。

参考文献:

[1] TAN W，LIU H，DONG Z，et al. Robust monocular SLAM in dynamic environments [C] // IEEE International Symposium on Mixed and Augmented Reality，IEEE，2013：209 - 218.

[2] WANG C C，THORPE C. Simultaneous localization and mapping with detection and tracking of moving objects [C] // Proceedings of 2002 IEEE International Conference on Robotics and Automation（ICRA），IEEE，2002：2918 - 2924.

[3] KLEIN G，MURRAY D. Parallel tracking and mapping forsmall AR workspaces [C] // The 6th IEEE and ACM International Symposium on Mixed and Augmented Reality，Nara：IEEE，2007：225 - 234.

[4] MUR-ARTAL R，MONTIEL J M M，TARDS J D. ORB-SLAM：A versatile and accurate monocular SLAM system [J]. IEEE Transactions on Robotics，2015，31（5）：1147 - 1163.

[5] MUR-ARTAL R，TARDOS J D. ORB-SLAM 2：An open-source SLAM system for monocular，stereo，and RGB-D cameras [J]. IEEE Transactions on Robotics，2016（99）：1 - 8.

[6] HE K，ZHANG X，REN S，et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2015，37（9）：1904 - 1916.

[7] GIRSHICK R. Fast R-CNN [C] // IEEE International Conference on Computer Vision，IEEE Computer Society，2015：1440 - 1448.

[8] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2015，39（6）：1137 - 1149.

[9] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection [C] // International Conf on Computer Vision and Pattern Recognition，IEEE，2016：779 - 788.

[10] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single shot multi-box detector [C] // European Conference on Computer Vision，Springer，Cham，2016：21 - 37.

[11] HOLSCHNEIDER M，KRONLAND-MARTINET R，MORLET J，et al. A real-time algorithm for signal analysis with the help of the wavelet transform [M]. Berlin Heidelberg：Springer，1989：286 - 297.

[12] RUBLEE E，RABAUD V，KONOLIGE K，et al. ORB：An efficient alternative to SIFT or SURF [C] // IEEE International Conference on Computer Vision，IEEE，2012：2564 - 2571.

[13] GAO X S，HOU X R，TANG J，et al. Complete solution classification for the perspective-three-point problem [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2003，25（8）：930 - 943.

[14] PENATE-SANCHEZ A，ANDRADE-CETTO J，MORENO-NOGUER F. Exhaustive linearization for robust camera pose and focal length estimation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2013，35（10）：2387.

[15] TRIGGS B，MCLAUCHLAN P F，HARTLEY R I，et al. Bundle adjustment——A modern synthesis [C] // International Workshop on Vision Algorithms：Theory and Practice，Springer-Verlag，1999：298 - 372.

[16] STURM J，ENGELHARD N，ENDRES F，et al. A benchmark for the evaluation of RGB-D SLAM systems [C] // IEEE/RSJ International Conference on Intelligent Robots and Systems，IEEE，2012.

（下转第 132 页）

3) 通过 ASM 算法得到的 3 个特征值作为 SVM 分类器的输入,分类得出该帧图像中驾驶员的人眼状态。

将第一部分测试样本分出 4 个不同的场景样本集进行测试。采用交叉验证的方式训练和测试 SVM。在测试样本集中,人为地标记出来每一帧图像人脸的眼睛状态,作为 SVM 分类器的训练和测试样本。最终测试的分类结果如表 1 所示,可以看出: 分类器较好地将人眼状态分类。

表 1 人眼状态分类结果

样本集	帧区间	闭眼帧数	睁眼帧数	分类准确度/%
174	7 485 ~ 7 636	12	140	90.79
164	4 612 ~ 4 667	16	40	85.71
150	2 477 ~ 2 539	7	56	86.48
86	2 022 ~ 2 115	13	81	82.61

改进的 ASM 算法通过计算标定点和最终获取到的关键点之间的平均欧氏距离作为衡量性能的标准,计算得到的欧氏距离作为对本文算法的评测标准

$$E=\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{n}\sum_{j=1}^{n}\left(\sqrt{\left(x_{ij}-x'_{ij}\right)^2+\left(y_{ij}-y'_{ij}\right)^2}\right)\right\} \quad (10)$$

式中 N 为测试图片数量,n 为单幅图片关键点数量。本文 N 取 50,n 取 20。改进前传统 ASM 算法平均误差为 42.51 像素,改进后 ASM 算法平均误差为 37.72 像素,性能提高了 11.2%。

5 结 论

实验结果表明: SVM 分类器可以将获取到的每一帧驾驶员人眼状态较好地进行分类,同时改进的 ASM 算法对于光线的变换具有较好的鲁棒性。

列车行驶过程中涉及到复杂的机车操作和瞭望等动作,驾驶员会频繁出现低头操作和张望动作,该算法针对此类场景还存在无法检测的情形。如何进一步排除真实环境下驾驶过程中噪声的干扰,依旧是一个需要研究的问题。

参考文献:

[1] 李都厚,刘群,袁伟,等. 疲劳驾驶与交通事故关系[J]. 交通运输工程学报,2010(2):104 - 109.

[2] 周凌霄. 多源生理信号融合的驾驶疲劳检测预警系统研究[D]. 杭州: 杭州电子科技大学,2015.

[3] 孙香梅. 基于车辆运行轨迹的疲劳驾驶检测研究[D]. 长沙: 长沙理工大学,2012.

[4] 黄皓. 基于驾驶操作及车辆状态的疲劳驾驶行为检测研究[D]. 南京: 东南大学,2016.

[5] 毛须伟,景文博,王晓曼,等. 一种基于眼部状态的疲劳驾驶检测方法[J]. 长春理工大学学报: 自然科学版,2016(2):125 - 130,136.

[6] 于兴玲,王民,张立材. 驾驶员眼睛疲劳状态检测技术研究[J]. 传感器与微系统,2007,26(7):16 - 17,20.

[7] 姚胜,李晓华,张卫华,等. 基于 LBP 的眼睛开闭检测方法[J]. 计算机应用研究,2015(6):1897 - 1901.

[8] 李杰. 可见光/近红外人脸识别方法的研究与实现[D]. 北京: 北京交通大学,2016.

[9] 陈宇波,陈新林,陈守明. 一种在红外图像中定位人眼的方法[J]. 传感器与微系统,2010,29(4):62 - 66.

[10] VAN GINNEKEN B,FRANGI A F,STAAL J J,et al. Active shape model segmentation with optimal features[J]. IEEE Transactions on Medical Imaging,2002,21(8):924 - 933.

[11] ZHAO Y,JIA W,HU R X,et al. Completed robust local binary pattern for texture classification[J]. Neurocomputing,2013,106:68 - 76.

[12] LIAO S,ZHU X,LEI Z,et al. Learning multi-scale block local binary patterns for face recognition[C]// 2007 Proceedings of the International Conference on Biometrics,Springer,2007.

[13] WANG J,ZHENG J,ZHANG S,et al. A face recognition system based on local binary patterns and support vector machine for home security service robot[C]// Proceedings of 2016 the 9th Internation al Symposium on Computational Intelligence and Design(ISCID),IEEE,2016.

[14] GAO W,CAO B,SHAN S,et al. The CAS-PEAL large-scale Chinese face database and baseline evaluations[J]. IEEE Transactions on Systems,Man,and Cybernetics——Part A: Systems and Humans,2008,38(1):149 - 161.

作者简介:

王 帅(1992 -),男,硕士研究生,研究方向为计算机视觉,嵌入式系统。

李凤荣,女,通讯作者,博士,助理研究员,研究领域为网络协议、网络安全,E-mail: fengrong. li@ hansuotech. com。

(上接第 128 页)

作者简介:

王金戈(1993 -),男,硕士研究生,研究方向为机器人环境感知与定位导航技术,E-mail: wjg172184@ 163. com。

邹旭东(1986 -),男,博士,研究员,主要研究领域为 MEMS 惯性传感技术、无人平台自主定位与导航技术、微系统集成技术。

仇晓松(1993 -),女,硕士研究生,研究方向为机器人环境感知与定位导航技术。

蔡浩原(1977 -),男,博士,副研究员,主要研究领域为智能微传感器。