实时的移动机器人语义地图构建系统*

李秀智[1,2],李尚宇[1,2],贾松敏[1,2],单吉超[1,2]
(1. 北京工业大学信息学部 北京 100124;2. 数字社区教育部工程研究中心 北京 100124)

摘 要:语义信息可以使机器人更充分地理解未知环境,为更高级的人机交互和完成更复杂的任务奠定基础。为了能够使移动机器人实时地创建语义地图,在 Jetson TX1 嵌入式电脑上开发了一种轻量级的深度学习目标检测模型,在保证了检测精度的同时,实现了高效的目标检测功能。并利用了视频流中的帧间光流信息,使用运动信息指导传播算法降低检测算法的漏检率。对于 Kinect 传感器生成的深度图像有黑边、黑洞等缺陷,使用统一计算设备架构(CUDA)技术开发了一种实时的深度图像修复算法。利用即时定位与地图构建(SLAM)技术,实现移动机器人底层的定位、导航、地图创建功能,并在此基础上使用贝叶斯推理框架,同时融合了环境的度量信息与视觉识别信息完成了语义地图的创建。经过实验表明,所提出的方法在实际的、复杂的室内环境下可以使移动机器人实时地创建语义地图。

关键词: 深度学习;图像修复;语义地图;贝叶斯推理;统一计算设备架构
中图分类号: TP242 TH72 **文献标识码**: A **国家标准学科分类代码**: 510.4050

System of real time mobile robot semantic map building

Li Xiuzhi[1,2], Li Shangyu[1,2], Jia Songmin[1,2], Shan Jichao[1,2]
(*1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*;
2. Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China)

Abstract: Semantic information can help the robot to better understand unknown environment and lay the foundation for more advanced human-computer interaction and more complicated task. To enable mobile robot to build semantic map in real time, a light deep learning model is developed for object detection on embedded computer Jetson TX1. The inter-frame optical flow information in the video stream is used to reduce the missing rate of object detection algorithm, which is called motion guided propagation (MGP) algorithm. A real-time depth map restoration algorithm based on CUDA is utilized because the depth map generated by Kinect has black hole and black border. SLAM technology is employed in this paper for robot location, navigation and mapping. On this basis, Bayesian inference framework is integrated with measurement information of environment and object detection information to complete the building of semantic map. Experiments show that the proposed method can enable the mobile robot to build the semantic map in real time in the real, complicated indoor environment.

Keywords: deep learning; image restoration; semantic map; Bayesian inference; compute unfied device architecture (CUDA)

0 引 言

随着机器人研究取得的卓越进展,可以预见在不远的未来机器人可以适应更加复杂的未知环境,可以实现更高级的人机交互,从而在日常生活中帮助或替代人们完成不同的任务,比如房屋清洁、安保、护理、娱乐等等[1]。

地图是编码环境信息的重要载体之一。过去十年来,关于如何表达、建立并维护地图成为了机器人研究中最热门领域之一[2]。传统的地图形式,例如栅格地图[3]和拓扑地图[4]可以满足机器人的像导航、定位、路径规划等基础功能。但是这些地图形式不包含环境的高层次语义信息,而这些信息对于机器人更充分地理解环境、执行更高级的人机交互任务都是至关重要的。比如栅格地图

收稿日期:2017-07 Received Date: 2017-07
* 基金项目:北京市教育委员会科技计划面上项目(KM201510005005)、北京工业大学智能制造领域大科研推进计划(040000546317552)项目资助

可以表达出一个房间的几何信息,但是并不能指示出房间中有什么物体,也不能描述房间的属性和功能。这便是本文研究机器人语义地图创建技术的动机,在未来智能家居、医疗助残等领域一定有广泛的应用价值。

国内外的许多研究机构、学者都投入到了机器人语义地图创建技术的研究中来,由于不同方法所使用的技术与要解决的问题不同,所以不同研究者对“语义地图”的定义与理解也不尽相同[5]。山东大学的吴皓等人[6-7]使用 QR code 技术,在家庭半未知环境下,对大物体粘贴二维码作为人工路标从而构建能描述物品-房间归属关系的语义地图;赵程[8]通过视觉跟踪人体与语音标注技术实现了一种自下而上的栅格-拓扑-语义多层次地图,但是在建图的过程中依赖于人工介入;文献[9]提出了通过建立三维室内语义地图实现房间识别从而帮助视力有障碍的人群;文献[10]实现了基于全局的条件随机场下的大规模道路场景致密语义地图构建;Sheng W 等人[11]创造性地提出了使用可穿戴设备来识别人体的动作,并建立了一个基于人体动作与物体种类关系的贝叶斯框架来构建语义地图,但可穿戴设备的佩戴对于实际应用略显繁琐。

无论是通过二维码[6-7]等人工标记,还是可穿戴设备[11]等辅助传感器,这些语义地图构建的方法都不够直接、简洁。此外,目前大多数关于语义地图的研究都是针对结构化环境[10]、场景简单的实验环境[11]或是仿真环境,对于基于真实的、非结构化生活场景的语义地图构建研究较少。其原因是传统的视觉检测方法计算效率低、检测鲁棒性差、模型泛化能力不强,难以仅通过视觉传感器实现移动机器人的实时语义地图构建。于金山等人[12]提出了基于云的语意库设计及语义地图构建,将复杂的视觉任务转移到云端计算。近年来,深度学习在目标检测领域取得了令人瞩目的进步,文献[13-16]中一系列基于卷积神经网络的目标检测方法使计算机视觉技术实用化、工程化成为了可能。但即使是目前领先水平的目标检测算法[16]在嵌入式系统上的执行效率也无法达到实时的要求。

本文在实验室机器人未知环境探索建图模块研究基础上[17],针对实际的复杂室内场景,提出了一种基于深度学习技术的移动机器人实时语义地图创建方法。首先利用统一计算设备架构(compute unified device architecture, CUDA)技术,克服了传统图像修复算法效率低的缺陷[18],针对 Kinect v2 获取的深度图像有黑边、黑洞的现象提出了一种实时的深度图像修复算法;其次优化了文献[16]的网络结构,训练了一个轻量级的目标检测模型,从而实现在嵌入式电脑上实时的目标检测,并利用了视频流中的帧间光流[19]信息,使用运动信息指导传播算法降低检测算法的漏检率;最后本文在贝叶斯推理框架[20]下,同时融合了环境的度量信息与视觉识别信息完成了语义地图的创建;实验证明本文方法在实际的室内环境下可以使移动机器人实时地创建语义地图。

1 移动机器人实时语义地图构建框架

本文所述语义地图构建框架如图 1 所示。硬件上将图像获取与预处理模块以及目标检测模块这两个适用于 CUDA 加速的机器视觉模块部署在 Jetson TX1 上[21],将机器人的基础控制、定位、建图、以及语义地图构建模块在研华工控机上使用机器人操作系统(robot operating system, ROS)管理执行,两块嵌入式电脑之间通过 SOCKET 进行通信。

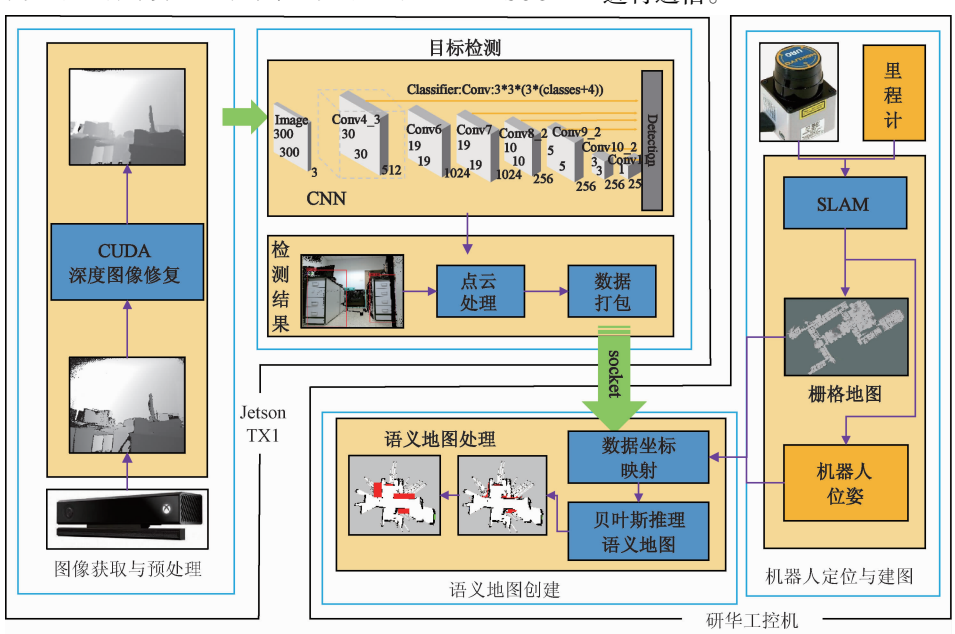


图 1 语义地图构建框架

Fig. 1 Framework of semantic mapping building

算法上,首先使用 Kinect v2 来同时获取场景的深度图像与彩色图像,然后对原始深度图像进行无效点的修复处理;接着使用卷积神经网络对彩色图像进行目标检测,并通过对应的深度图像计算出物体在摄像机坐标系下的三维坐标 (x,y,z) 和识别出物体的类别信息 l 一同打包传输到机器人主控电脑;在工控机上结合即时定位与地图构建(simultaneous localization and mapping, SLAM)算法输出的机器人定位信息以及接收到的识别数据进行识别物体的坐标映射,统一到地图坐标系下;最后,使用贝叶斯推理增量式地构建语义地图。

2　基于 CUDA 的实时深度图像修复算法

深度图像可以直接反应出真实的三维环境信息,如图 2 所示,由于遮挡、反射以及光线剧烈变化等原因,Kinect 生成的原始深度图像存在着黑边、黑洞等大量的无效区域,这对深度图像的使用造成了很大的影响。为此,本文提出了一种使用 CUDA 技术的并行化实时深度图像修复算法,以实现在移动机器人上实时有效地修复深度图像。

图 2　原始深度图像

Fig. 2　Raw depth image

为了使图像修复程序并行化,首先要对图像进行划分。Kinect v2 的深度图像大小为 512 ×424,略去图像上下各 12 行的像素,以 32 ×20 为一个 block,构成 16 ×20 的 grid。图像划分完成后便上传到 GPU 并行执行图像修复程序。对每一个图像上的无效点使用式(1)进行滤波。

$$I_{\text{dest}}(x,y_{(x,y)\in\Omega_{\text{inv}}}) = \frac{1}{\omega_p}\sum_{i,j\in\Omega_n}\omega(i,j)\cdot I_{\text{src}}(i,j) \tag{1}$$

式中:I_{dest} 是修复后的图像,I_{src} 为原图像,$\omega(i,j)$ 为滤波器在点(i,j)的权,Ω_{inv} 为图像上的无效点区域,Ω_n 是除去无效点的像素邻域,ω_p 是标准量由式(2)计算。

$$\omega_p = \sum_{i,j\in\Omega_n}\omega(i,j) \tag{2}$$

而权值 $\omega(i,j)$ 同时与像素点的空域与值域线性相关,距离越近、像素值变化越小相关性越大,其滤波核函数定义如下:

$$\omega(i,j) = \exp\left(-\frac{(i-x)^2+(j-y)^2}{2\partial_d}\right)\cdot \exp\left(-\frac{(I(i,j)^2-I(x,y))^2}{2\partial_I}\right) \tag{3}$$

在 GPU 端执行的程序称为核函数,对于本文方法同时启动了 20 ×32 ×20 ×16 =204 800 个线程,204 800 个核函数同时修复深度图像相比于在 CPU 上顺序执行,效率提升明显。

3　基于轻量级卷积神经网络的目标检测

文献[16]提出的单次输入多包围框检测器(single shot multibox detector, SSD)方法是一种基于卷积神经网络(convolutional neural networks, CNN)的端到端的目标检测模型,它同时输出一组固定尺寸的检测框和和它们的分类置信度,SSD 框架如图 3 所示。在 NVIDIA Titan X 上,对于 300 ×300 大小的输入图像,在 VOC2007 test 数据集上 SSD 达到了 72.1 的平均准确率(mean average precision, mAP),速度为每秒 58 帧(frames per second, FPS)。

尽管文献[16]中的方法在高性能 GPU 上的速度已经达到了 58 FPS,但经过实验发现即使是在 TX1 这款针对深度学习设计的高性能嵌入式电脑上,原始的 SSD 模型的检测速度只有 1 ~2 FPS,显然无法满足移动机器人实时性的要求。

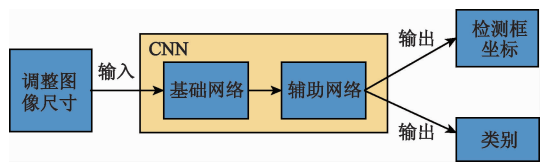


图 3　SSD 框架

Fig. 3　Framework of SSD

3.1　模型结构

SSD 的主体结构分为基础网络与辅助网络两部分,基础网络是由高质量图像分类的标准架构截断而成,而辅助网络添加到基础网络的末尾完成了多尺度特征图的检测,最后同时输出了检测框的坐标以及类别。本文注意到原始的 SSD 模型在基础网络上采用了文献[22]中的 VGG-16 模型,该模型有着强大的图像分类能力,但是如图 4(a)所示,16 层的网络深度带来的是近 1.38×10^8 个参数,巨大规模的参数量必然会导致算法的效率低下。考虑到本文的目的主要是针对室内环境的固定的、标志性物体进行识别,而且对算法在嵌入式系统执行的实时性要求很高,本文提出了一种轻量级的目标检测模型来替代 VGG-16 网络作为基础网络,称之为 Light-L 网络。如图 4(b)所示,Light-L 模型的结构也是由输入层-卷积层-池化层,这样经典的结构组合而成,但是网络的深度与卷积核的个数都大幅下降,该模型的参数约有 1.5 ×

10^6个，比 VGG-16 减少了约 1.36×10^8 个。实验证明本文提出的轻量卷积神经网络模型，在保证识别率在可以应用的前提之下效率有了大幅提升，使在移动机器人的嵌入式系统上执行深度学习算法成为了可能。

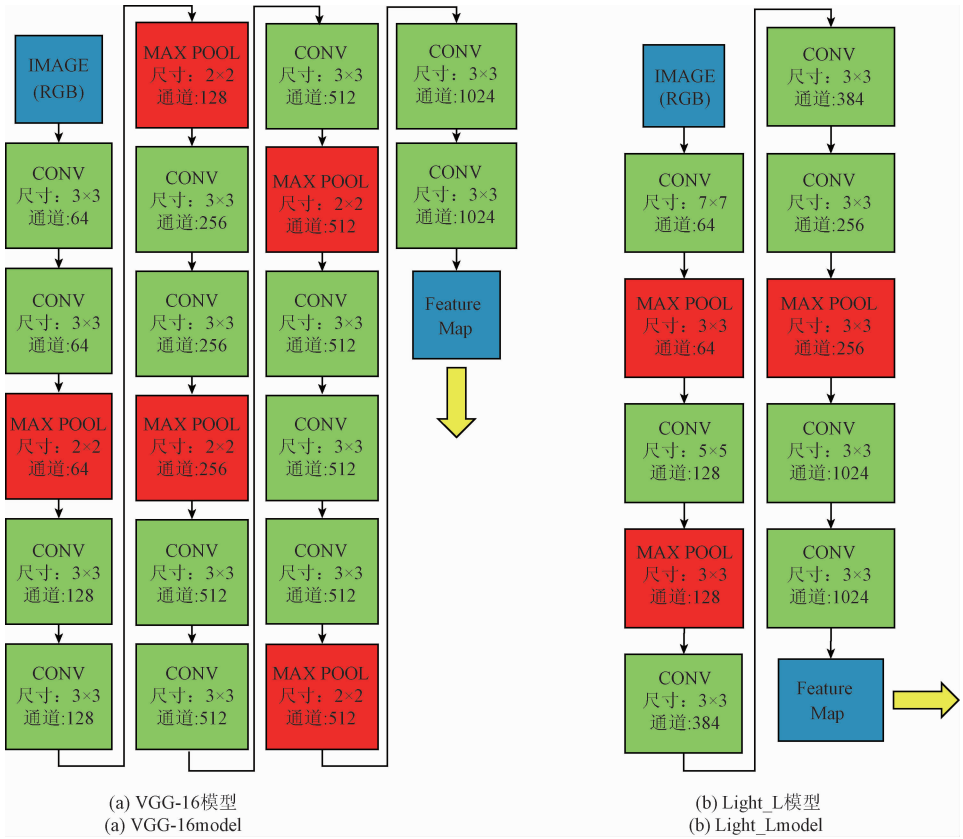


图 4 基础网络结构

Fig. 4 Architecture of base network

3.2 网络训练

ImageNet、VOC 等图像数据库为研究者们提供了海量的带有标注信息的图像数据，极大地方便了各种图像分类、检测模型的训练与验证。然而这些数据库中的图片大多是以人类的视角拍摄的，本文使用的机器人视角较低，对于同样种类的物体，比如“沙发”、“办公桌”，获取的图片与人类视角拍摄图片差距较大，为了使机器人更有针对性地理解室内环境，作者所在团队手工标注了近 5 000 张机器人视角下的图片来构成训练卷积神经网络模型的数据库。

本文采用文献[16]中的模型训练方法，训练的目标损失函数为：

$$L(x_{ij}^p, c, l, g) = \frac{1}{N}(L_{\text{conf}}(x_{ij}^p, c) + \alpha L_{\text{loc}}(x_{ij}^p, l, g)) \tag{4}$$

式中：$x_{ij}^p = 1$ 表示第 i 个默认框匹配上了第 j 个真值框，属于第 p 个类别；N 为匹配的默认框总数；L_{conf}为分类的置信损失，是对于多类别置信度 c 的 softmax 损失函数；L_{loc}为位置损失，是预测框 l 和真实标签值框 g 参数之间的 L1 Smooth 损失函数[14]。

3.3 运动指导传播

较差的拍摄角度或相机运动造成的图像模糊等原因可能会导致目标检测算法的漏检，由于视频流中相邻帧的检测结果无论是位置还是检测置信度都应该是高度相关的，这些漏检物体可以被相邻帧间的检测信息恢复出来。这启发本文将检测框和它的置信度传播到相邻帧从而提高检测算法的召回率。

如图 5 所示，首先对视频中一帧图像通过上述卷积神经网络进行静态图像检测；而后利用光流信息[24]将每一个检测框传播到下一帧图像；最后使用非极大抑制方法去除冗余的检测框。表 1 所示为该算法的流程。

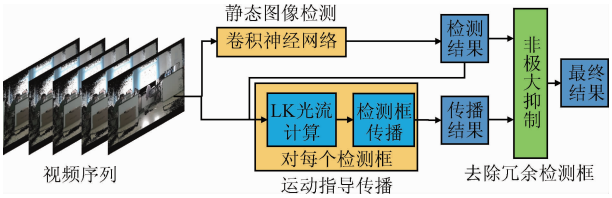


图 5　运动指导传播框架

Fig. 5　Framework of motion guided propagation

表 1　运动指导传播算法

Table 1　The motion guided propagation algorithm

算法 1 运动指导传播(motion guided propagation, MGP)

输入:第 t 帧的检测结果//每一个检测框的坐标、类别与检测分数

输入:t 与 $t+1$ 帧两幅图像

1:**for each** B_t^i **in** B_t do//B_t 为 t 帧上所有未遍历的检测框

2:　将 B_t^i 设为图像 ROI//B_t^i 为第 t 帧图像上的第 i 个检测框

3:　提取图像角点 points[t]

4:　使用 Lucas-Kanade[23] 光流算法跟踪角点 points[$t+1$]

5:　**if** points[$t+1$].size/points[t].size < *threshold* **then**//跟踪失效点过多

6:　　**continue**

7:　**end if**

8:　计算 **mean_OF**//平均光流矢量

9:　计算传播后的检测框位置

10:置信度传播 B_t^i confidence→B_{t+1}^i confidence

11:**end for**

3.4　数据打包与传输

对于每一帧机器人观测到的图像,由上述目标检测算法输出的结果为检测到目标的类别 l 与检测框(x, y, w, h)。其中(x, y)是检测框的左上角图像坐标,w 和 h 分别为检测框的宽度和高度,单位是像素。由于 Kinect 可以直接感知环境的深度信息,使用 libfreenect2 库的 RGB 图像与深度图像配准函数便可以计算出 RGB 图像中检测框所包围的点对应的三维点云。

本文将每一个检测框包围的三维点云中的所有有效点的重心坐标作为构建语义地图的检测点的输入:

$$(x_c, y_c, z_c) = \left(\frac{\sum_{i \in S_p} x_i}{n}, \frac{\sum_{i \in S_p} y_i}{n}, \frac{\sum_{i \in S_p} z_i}{n} \right) \tag{5}$$

式中:(x_c, y_c, z_c) 为在摄像机坐标系下被检测物体的空间坐标,S_p 为 RGB 图像检测框中包含点对应的三维点云中的所有有效点集合,n 是该检测框包围的点云中有效点的个数。最后将数据打包成 <S,l, x_c, y_c, z_c > 的元组形式通过 SOCKET 通信发送给机器人主控端。其中 S 为数据标志位,目的是便于数据解码,保证通信的鲁棒性。

4　语义地图构建

为了使机器人能够理解环境的语义信息从而执行更高级的人机交互任务,本文在机器人未知环境探索、2D 栅格地图构建[17] 等模块的研究基础之上提出了一种语义地图的构建方法。首先将识别到物体的坐标映射到栅格地图坐标系,继而使用贝叶斯推理技术结合检测点的几何度量信息与识别信息增量式地构建语义地图,最后对语义地图进行优化和存储。

4.1　坐标映射

在解码由 TX1 传输的检测信息后,得到了检测物体的类别 l 以及位置信息(x_c, y_c, z_c),需要将摄像机坐标系下的位置信息映射到栅格地图坐标系下。机器人坐标系定义如图 6 所示。

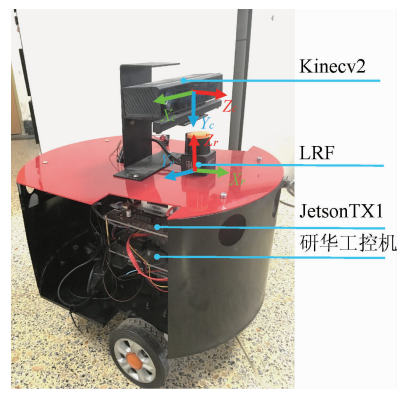


图 6　机器人坐标系定义

Fig. 6　Definition of robot coordinate system

机器人坐标系与摄像机坐标系的转换关系为:

$$\begin{bmatrix} X_r \\ Y_r \\ Z_r \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{cr} & \boldsymbol{T}_{cr} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{6}$$

式中:$\boldsymbol{R}_{cr}$ 与 $\boldsymbol{T}_{cr}$ 分别是机器人坐标系与摄像机坐标系之间的旋转矩阵和平移矩阵,由传感器的安装位置决定,对于本文的机器人有:

$$\boldsymbol{R}_{cr} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \boldsymbol{T}_{cr} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{7}$$

式中:t_x、t_y、t_z 是在摄像机坐标系下机器人坐标系的原点坐标,可由机器人安装尺寸或测量获得。机器人坐标系与地图物理坐标系(世界坐标系)的关系如下:

$$\begin{bmatrix} X_m \\ Y_m \\ Z_m \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{mr} & \boldsymbol{T}_{mr} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_r \\ Y_r \\ Z_r \\ 1 \end{bmatrix} \tag{8}$$

式中：$\boldsymbol{R}_{mr}$、$\boldsymbol{T}_{mr}$ 为世界坐标系与机器人坐标系之间的旋转平移矩阵，由 SLAM 算法获取。对于本文有：

$$\boldsymbol{R}_{mr} = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{T}_{mr} = \begin{bmatrix} t_{mx} \\ t_{my} \\ 0 \end{bmatrix} \tag{9}$$

式中：(t_{mx}, t_{my}, ϕ) 是由 2D SLAM 算法输出的机器人位姿。最后将世界坐标系下的坐标转换成离散的栅格坐标：

$$\begin{bmatrix} X_g \\ Y_g \end{bmatrix} = \mathrm{int}\left(\begin{bmatrix} X_m \\ Y_m \end{bmatrix} / resolution \right) \tag{10}$$

式中：*resolution* 是栅格地图的分辨率，int 表示取整函数。经过一系列的转换，得到了被检测到的物体在栅格坐标系下的坐标 (X_g, Y_g)。

4.2 贝叶斯推理增量式构建语义地图

尽管现有的深度学习模型在目标检测上已经取得了很高的精度，但是由于光照变化强烈、机器人运动速度过快等原因会造成目标检测算法的漏检、误检。因此本文定义了概率分布函数 $P_{G,t}(L)$ 表示在 t 时刻栅格 G 存在检测物体的概率，L 表示存在检测物体，继而便可使用贝叶斯推理来不断地更新检测到物体的概率。使用 $P(L|O)$ 来表示在算法检测到物体的情况下物体确实存在的概率，O 表示检测到物体。另外，本文注意到某一位置是否存在检测物体与这一位置对应的几何度量信息相关，而该信息可以通过底层的 2D 栅格地图获得。如图 7 所示，使用 M(G) 来表示栅格 G 的几何类别，$M(G) \in \{M1, M2, M3, M4\}$，$M1$ 表示检测点在栅格地图的障碍物区域；$M2$ 表示检测点在栅格地图的未知区域；$M3$ 表示检测点在栅格地图的自由区域但附近有障碍物；$M4$ 表示检测点在栅格地图的自由区域而附近也没有障碍物，其中定义点的附近区域要根据具体的环境来设置距离阈值，比如 20 cm。栅格的几何信息与该栅格内存在检测物体的概率相关性可由先验知识 $P(M(G)|L)$ 表示。

图 7 栅格类别定义

Fig. 7 Definition of grid type

有了卷积神经网络的检测结果以及栅格所属的几何类别这两个通道的信息，便可以使用贝叶斯推理更新一个栅格存在被检测物体的后验概率：

$$P_{G,t}(L|O_t, M(G)) = \frac{P(O_t, M(G)|L) P_{G,t-1}(L)}{P(O_t, M(G))} \tag{11}$$

由于两个通道的信息是独立的，可以推导出：

$$P_{G,t}(L|O_t, M(G)) = \frac{P(O_t|L) P(M(G)|L) P_{G,t-1}(L)}{P(O_t) P(M(G))} \tag{12}$$

式中：$P_{G,t-1}(L)$ 是基于 $t-1$ 时刻之前所有时间内的关于栅格 G 上的检测物体信息的先验知识。可以将后验概率 $P_{G,t}(L|O_t, M(G))$ 简写成 $P_{G,t}(L)$，作为下一次迭代更新的先验概率。于是，在 t 时刻，某一栅格 G 内发生了事件 O，也就是检测到了物体，便利用式(12)对该栅格的概率值进行更新。同样的，对于逆事件 $\bar{O}$，利用式(13)进行更新。

$$P_{G,t}(L|\bar{O}_t, M(G)) = \frac{P(\bar{O}_t|L) P(M(G)|L) P_{G,t-1}(L)}{P(\bar{O}_t) P(M(G))} \tag{13}$$

最后，在机器人探索结束时刻，通过栅格 G 的最终后验概率是否大于设定的阈值来判定该栅格是否存在被检测物体。

4.3 语义地图处理与存储

由上述方法构建的语义地图(见图 7)，是由一系列孤立的标注点来表示在某一位置存在某种类型的物体。为了更好地表征环境的高层次语义信息，对构建的原始语义地图进行优化处理，使用彩色矩形块来表示环境中某一区域存在某种类型的物体。

首先对原始语义地图上不同类别的标注信息看作图像，对之进行形态学闭运算，这样便可连接邻近的同类标注点；然后对地图做联通区域分析，去除过小的标注区域；最后，对保留下来的标注区域取外接矩形。这样就得到了相对简洁、清晰的环境语义地图。

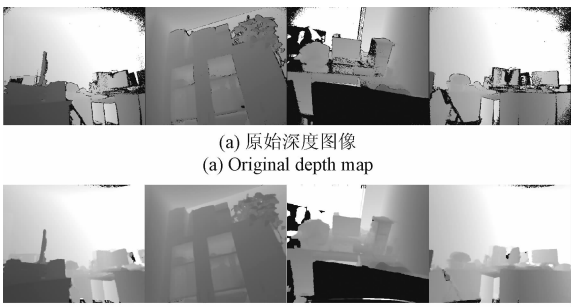
除了以图片的形式表示语义地图，本文还建立了语义地图描述文件来存储语义地图信息。将每一个语义地图上的物体都以元组 $\{id, label, x, y, w, h\}$ 的形式存储在文本文件中，其中 id 为物体的编号，$label$ 为物体的类别代码，(x, y) 表示物体在栅格地图坐标系下的坐标，w、h 分别表示该物体在处理后的语义地图上矩形块的宽度和高度，此外也将栅格地图与真实环境的尺度信息 *resolution* 记录在文本中。有了上述语义地图描述文件，便可以实现标注物品的真实位置查询，从而方便机器人执行更高级的任务。

5　实验结果与分析

实验硬件为作者所在团队自主研发的移动机器人平台(见图 6)。机器人主控单元为研华工控机,CPU 为 Intel Core i7,主频 2.5 GHz,RAM 为 4 GB。另外机器人还配备一块 Jetson TX1 嵌入式电脑负责视觉算法的执行,它拥有 256 个 NVIDIA CUDA 核心和 64 位 CPU。传感器方面,机器人配有 URG-04LX-UG01 激光测距仪和 Kinect v2 深度相机。

5.1　深度图像修复实验

为了验证算法有效性,本文选取了 4 幅在实验室环境下拍摄的真实深度图像以及经过处理后的图像作为比较,结果如图 8 所示。图 8(a)所示为原始的深度图像,图 8(b)所示为经过本文的图像修复算法处理后的深度图像,可以看到原始图像中出现的黑边、空洞等无效点可以被有效地修复。



(a) 原始深度图像
(a) Original depth map

(b) 处理后的深度图像
(b) The depth map after processig

图 8　深度图像修复算法结果

Fig. 8　Results of proposed depth map restoration method

为了验证本文使用 CUDA 技术对深度图像修复算法执行效率的提升效果,本文分别在一款配备 GTX1070 高性能显卡的台式机以及 Jetson TX1 嵌入式电脑两个平台上分别统计了算法在 CPU 与 GPU 下的执行效率。其中,卷积核的滤波框大小为 11 个像素,FPS 的单位是帧,即每秒钟可以处理图像的帧数。具体的结果如表 2 所示,台式机上加速比为 35.33,在 TX1 上加速比为 45,可以看出 CUDA 技术显著地提高了算法执行的效率。

表 2　深度图像修复算法在 CPU 与 GPU 下效率对比

Table 2　Efficiency comparison for deep map restoration algorithm in CPU and GPU　(帧)

	CPU i7-6700	GPU GTX1070	CPU TX 1	GPU TX 1
FPS	15	530	2	90

5.2　目标检测算法实验

本文目标检测模型的网络结构,由图 4(b)中的 Light_L 网络作为基础网络,后面连接的是与文献[16]相同的辅助网络。本文在自己建立的机器人视角下的数据库重新训练了网络,训练方法为随机梯度下降(stochastic gradient descent, SGD),初始学习率为 0.001,动量为 0.9,权重衰减为 0.000 5,批量尺寸为 16。

本文提出的物体检测算法的结果如图 9 所示,可以看出即使是在相对复杂、狭窄的实验室环境,在机器人运动过程中以及低矮的视角下本文使用的物体检测算法仍然表现良好。

图 9　目标检测算法结果

Fig. 9　Results of proposed object detection algorithm

表 3 所示为不同的基础网络下,在本文使用的数据库中,目标检测算法的效率与精度对比。其中第 1 行是原始 SSD 算法使用的基础网络,第 2 行是使用本文提出的 Light_L 模型作为基础网络的结果。可以看出在 GTX1070 下,算法的 FPS 由 49 提高到了 140,加速比是 2.857,在 Jetson TX1 下算法的 FPS 由 1.5 提高到了 8,加速比是 5.3,已经可以使机器人实时地运行该检测算法。此外,随着检测效率的大幅提升,算法的检测精度只有略微的下降,由 94.2% 下降到了 91.6%,满足本文语义地图创建算法的需要。

表 3　不同基础网络下检测算法效率与精度对比

Table 3　Comparison of efficiency and precision for CNN object detection network with different base nets

硬件平台	基础网络	FPS	平均准确率
GTX1070	VGG_16	49	0.942
	Light_L	140	0.916
Jetson TX1	VGG_16	1.5	0.942
	Light_L	8	0.916

5.3 运动指导传播算法实验

图 10 定性地展示了运动指导传播算法的效果，本文选取了一段视频中的连续 5 帧作为实验对象。如第 1 行所示 2 513 和 2 514 两帧都漏检了中间的桌子，而他们的相邻帧没有出现漏检。如图 10 中传播后图像所示，通过运动指导传播算法，漏检的桌子被成功恢复。

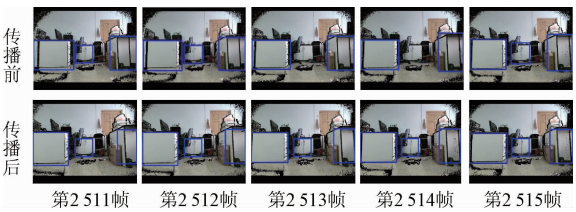


图 10 运动指导传播算法效果

Fig. 10 Effect of motion guided propagation algorithm

为了定量地度量该算法对漏检的抑制效果，选取了 4 段机器人在运动过程中拍摄的视频进行实验，通过指标召回率来评价检测算法的漏检情况，如表 4 所示。其中：

$$recall = \frac{TP}{TP + FN} \tag{14}$$

式中：TP 表示真阳例，FN 表示假阴例，召回率越高漏检率便越低。从表 4 中可以看出，运动指导传播算法可以有效的提升检测算法的召回率。其中在第 3 段视频中，检测的召回率较低，这是由于视频中的待检测物体与机器人距离较远。但即使是在这种极端的情况下，该算法仍将召回率提升了约 20%，再次验证了视频流中上下文信息确实可以弥补静态单帧图像检测的缺陷。

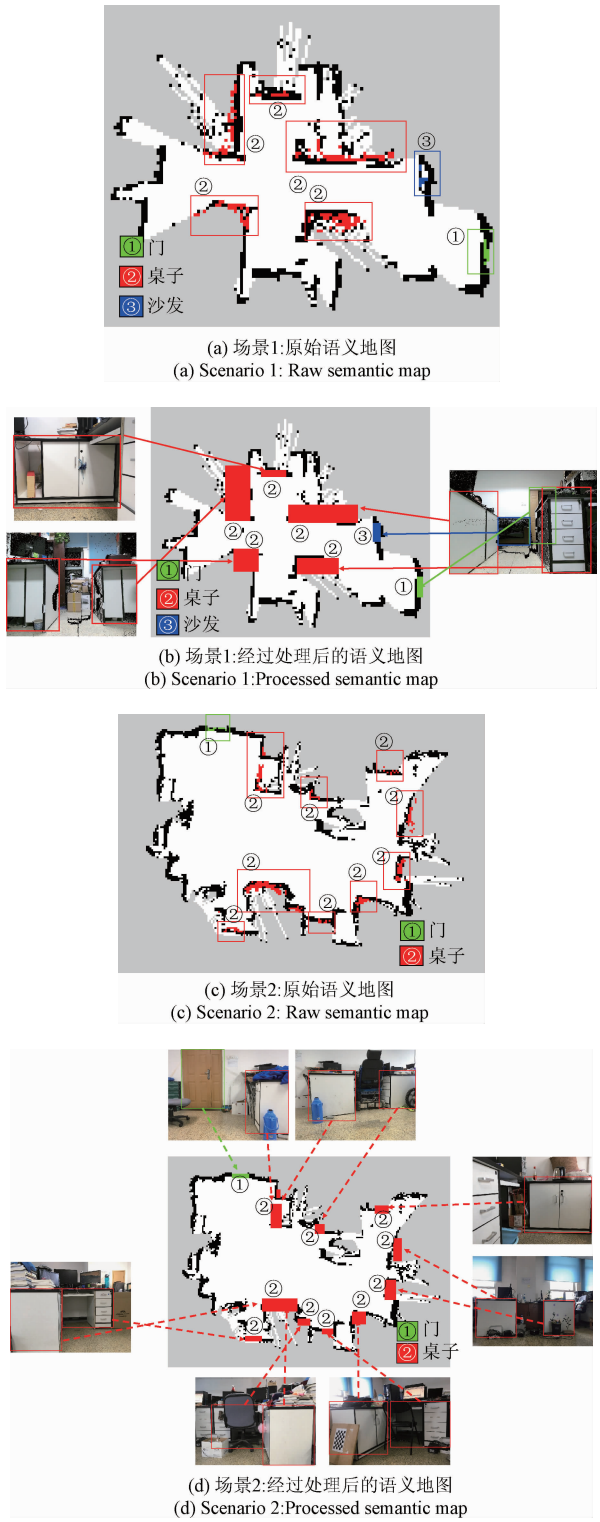
表 4 运动指导传播算法对识别召回率的影响

Table 4 Performance of MGP on detection recall

数据	是否使用 MGP 算法	召回率
视频 1	×	0.85
	√	0.88
视频 2	×	0.84
	√	0.91
视频 3	×	0.34
	√	0.53
视频 4	×	0.60
	√	0.69

5.4 语义地图创建实验

为了验证本文提出的语义地图构建方法，本文在实际场景下进行了实验。场景 1 是一间相对狭窄的实验室，机器人在该场景下构建的语义地图如图 11(a)所示。经过处理后的语义地图如图 11(b)所示。语义地图建立在栅格地图的基础上，栅格地图反映了环境结构的几何信息，地图中的黑色区域表示环境中的障碍物，白色区域表示自由区域，灰色区域表示未知区域，而编号①、编号②、编号③的彩色标注点表示该区域存在检测算法识别出的物体。场景 2 是一间包含更多物体的实验室，整体环境更加杂乱，在该场景下构建的语义地图如图 11(c)所示，经过处理后的语义地图如图 11(d)所示。本文机器人执行构建语义地图所在的实验场景均未经过人为的布置与整理，完全是日常实际使用的生活场景。将实验结果与真实环境比较，基于本文框架所构建的语义地图正确地反映了环境信息。



(a) 场景1:原始语义地图
(a) Scenario 1: Raw semantic map

(b) 场景1:经过处理后的语义地图
(b) Scenario 1:Processed semantic map

(c) 场景2:原始语义地图
(c) Scenario 2: Raw semantic map

(d) 场景2:经过处理后的语义地图
(d) Scenario 2:Processed semantic map

图 11 语义地图构建结果

Fig. 11 Results of semantic map building

6 结　论

本文提出了一个移动机器人实时语义地图构建系统。在硬件方面,本文研发了一款搭载研华工控机与 Jetson TX1 并配备有 Kinect v2、激光雷达等传感器的高性能移动机器人。在软件方面,使用 CUDA 技术实现了实时的深度图像修复。构建了 Light_L 轻量级卷积神经网络模型,来替换 SSD 方法中原始的基础网络,使得物体检测算法效率提升。并利用帧间光流信息,使用运动指导传播算法抑制了视频中图像检测的漏检率。本文在 2D SLAM 导航、定位、构建栅格地图算法的基础之上使用贝叶斯推理方法同时融合了环境的度量信息与视觉识别信息完成了语义地图的创建。实验验证了本文提出的语义地图构建系统可以使移动机器人在实际的、复杂的环境实时地构建语义地图。

参考文献

[1] WENG Y H, CHEN C H, SUN C T. Toward the human-robot co-existence society: On safety intelligence for next generation robots [J]. International Journal of Social Robotics, 2009, 1(4): 267-282.

[2] THRUN S. Robotic mapping: A survey [M]. San Francisco: Morgan Kaufmann Publishers Inc., 2002.

[3] 唐宏伟,孙炜,杨凯,等. 基于 SURF 特征的多机器人栅格地图拼接方法[J]. 电子测量与仪器学报, 2017, 31(6): 859-868.
TANG H W, SUN W, YANG K, et al. Grid map merging approach of multi-robot based on SURF feature[J]. Journal of Electronic Measurement and Instrumentation, 2017, 31(6): 859-868.

[4] SCHWERTHEGER S, BIRK A. Map evaluation using matched topology graphs[J]. Autonomous Robots, 2016, 40(5): 761-787.

[5] 朱博,高翔,赵燕喃. 机器人室内语义建图中的 w 场所感知方法综述[J]. 自动化学报, 2017, 43(4): 493-508.
ZHU B, GAO X, ZHAO Y N. Place perception for robot indoor semantic mapping: A survey[J]. Acta Automatica Sinica, 2017, 43(4): 493-508.

[6] 吴皓,田国会,薛英花,等. 基于 QR code 技术的家庭半未知环境语义地图构建[J]. 模式识别与人工智能, 2010, 23(4): 464-470.
WU H, TIAN G H, XUE Y H, et al. QR code based semantic map building in domestic semi-unknown environment [J]. Pattern Recognition and Artificial Intelligence, 2010, 23(4): 464-470.

[7] 吴皓,田国会,王家超,等. 室内非结构化环境三维栅格语义地图的构建[J]. 模式识别与人工智能, 2012, 25(4): 564-572.
WU H, TIAN G H, WANG J CH, et al. Three-dimensional grid semantic map building in unstructured indoor environment [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(4): 564-572.

[8] 赵程. 基于视觉—语音交互式室内层次地图构建与导航系统[D]. 厦门:厦门大学, 2014.
ZHAO CH. Indoor three layers mapping and navigation system based on visual-voice interaction [D]. Xiamen: Xiamen University, 2014.

[9] LIU Q, LI R, HU H, et al. Using semantic maps for room recognition to aid visually impaired people [C]. International Conference on Automation and Computing, 2016: 89-94.

[10] 江文婷. 大规模道路场景致密语义地图构建[D]. 杭州:浙江大学, 2016.
JIANG W T. Large scale road scene dense semantic mapping[D]. Hangzhou: Zhejiang University, 2016.

[11] SHENG W, DU J, CHENG Q, et al. Robot semantic mapping through human activity recognition: A wearable sensing and computing approach [J]. Robotics and Autonomous Systems, 2015, 68(C): 47-58.

[12] 于金山,吴皓,田国会,等. 基于云的语义库设计及机器人语义地图构建[J]. 机器人, 2016, 38(4): 410-419.
YU J SH, WU H, TIAN G H, et al. Semantic database design and semantic map construction of robots based on the cloud[J]. Robot, 2016, 38(4): 410-419.

[13] GIRSHICK R. Fast R-CNN [C]. IEEE International Conference on Computer Vision (ICCV), 2015: 1440-1448.

[14] 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.

[15] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.

[16] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector [M]. Berlin: Springer International Publishing, 2016: 21-37.

[17] LI X Z, QIU H, JIA S M, et al. Dynamic algorithm for safe and reachable frontier point generation for robot exploration [C]. IEEE International Conference on Mechatronics and Automation, 2016: 2088-2093.

[18] 李民,李世华,乐翔,等. 基于学习字典的图像修复算

法[J]. 仪器仪表学报, 2011,32(9): 2041-2048.
LI M, LI SH H, LE X, et al. Image inpainting algorithm based on learned dictionary [J]. Chinese Journal of Scientific Instrument, 2011,32(9): 2041-2048.

[19] 张聪炫,陈震,黎明. 金字塔光流三维运动估计与深度重建直接方法[J]. 仪器仪表学报, 2015,36(5): 1093-1105.
ZHANG C X, CHEN ZH, LI M. Direct method for 3D motion estimation and depth reconstruction based on pyramid optical flow [J]. Chinese Journal of Scientific Instrument, 2015,36(5): 1093-1105.

[20] 王滨,吕东辉. 基于贝叶斯判决的关于 YCbCr 空间的肤色模型查询表建立的研究[J]. 仪器仪表学报, 2004, 25(增刊2): 231-234.
WANG B, LV D H. A study of building skin color model of lookup table in YCbCr color space based on bayes decision [J]. Chinese Journal of Scientific Instrument, 2004, 25(Suppl. 2): 231-234.

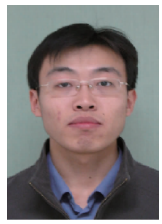
[21] WANG C, WANG Y, HANG Y H, et al. CNN-based object detection solutions for embedded heterogeneous multicore SoCs[C]. 22nd Asia and South Pacific Design Automation Conference (ASP-DAC), 2017: 105-110.

[22] CHEN L C, PAPABDREOU G, KOKKONOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs [J]. Computer Science, 2016(4): 357-361.

[23] LUCAS B D, KANADE T. An iterative image registration technique with an application to stereo vision [C]. International Joint Conference on Artificial Intelligence, 1981: 674-679.

[24] 伍济钢,王刚,蒋勉,等. 光流点匹配跟踪的薄壁件振动模态测试方法[J]. 电子测量与仪器学报, 2017, 31(6): 850-858.
WU J G, WANG G, JIANG M, et al. Vibration modal measurement method for thin-walled parts using optical flow point matching and tracking [J]. Journal of Electronic Measurement and Instrumentation, 2017, 31(6): 850-858.

作者简介



李秀智,2008 年于北京航天航空大学获得博士学位,现为北京工业大学副教授、硕士生导师,主要研究方向是智能机器人导航、机器视觉。
E-mail: xiuzhi. lee@163. com

Li Xiuzhi received his Ph. D. from the Beihang University in 2008. Now he is an associate professor and master supervisor in Beijing University of Technology. His main research interests include intelligent robot navigation and computer vision.

李尚宇(通讯作者), 2015 年于天津工业大学获得学士学位,现为北京工业大学在读硕士研究生,主要研究方向是机器视觉、计算机视觉。
E-mail: lishangyu1993@ hotmail. com

Li Shangyu (Corresponding author) received his B. Sc. degree from Tianjin Polytechnic University in 2015. Now, he is a M. Sc. candidate in Beijing University of Technology. His main research interests include machine vision and computer vision.