计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190578
57(6):1218-1226,2020

基于多视角 RGB-D 图像帧数据融合的室内场景理解

李祥攀[1] 张 彪[1] 孙凤池[2] 刘 杰[3]
[1](南开大学计算机学院 天津 300750)
[2](南开大学软件学院 天津 300750)
[3](南开大学人工智能学院 天津 300750)
(xiangpan.li@qq.com)

Indoor Scene Understanding by Fusing Multi-View RGB-D Image Frames

Li Xiangpan[1], Zhang Biao[1], Sun Fengchi[2], and Liu Jie[3]
[1](*College of Computer Science*, *Nankai University*, *Tianjin* 300750)
[2](*College of Software*, *Nankai University*, *Tianjin* 300750)
[3](*College of Artificial Intelligence*, *Nankai University*, *Tianjin* 300750)

Abstract For intelligent robots, it's an important and challenging ability to understand environment correctly, and so, scene understanding becomes a key problem in robotics community. In the future, more and more families will have service robots living with them. Family robots need to sense and understand surrounding environment reliably in an autonomous way, depending on their on-board sensors and scene understanding algorithms. Specifically, a running robot has to recognize various objects and the relations between them to autonomously implement tasks and perform intelligent man-robot interaction. Usually, RGB-D(RGB depth) visual sensors commonly used by robots to capture color and depth information have limited field of view, and so it is often difficult to directly get the single image of the whole scene in large-scale indoor spaces. Fortunately, robots can move to different locations and get more RGB-D images from multiple perspectives which can cover the whole scene in total. In this situation, we propose an indoor scene understanding algorithm based on information fusion of multi-view RGB-D images. This algorithm detects objects and extracts object relationship on single RGB-D image, then detects instance-level objects on multiple RGB-D image frames, and constructs object relation oriented topological map as the model of the whole scene. By dividing the RGB-D images into cells, then extracting color histogram features from the cells, we manage to find and associate the same objects in different frames using the object instance detection algorithm based on the longest common subsequence, overcoming the adverse influence on image fusion caused by RGB-D camera's viewpoint changes. Finally, the experimental results on the NYUv2 dataset demonstrate the effectiveness of the proposed algorithm.

Key words object detection; object instance detection; RGB-D image; object-relation topological map; scene understanding

收稿日期:2019-08-19;**修回日期**:2019-12-17
基金项目:国家自然科学基金项目(61873327)
This work was supported by the National Natural Science Foundation of China (61873327).
通信作者:孙凤池(fengchisun@nankai.edu.cn)

摘 要 对于智能机器人来说,正确地理解环境是一项非常重要且充满挑战性的能力,从而成为机器人学领域一个关键问题.随着服务机器人进入家庭成为趋势,让机器人能够依靠自身搭载的传感器和场景理解算法,以自主、可靠的方式感知并理解其所处的环境,识别环境中的各类物体及其相互关系,并建立环境模型,成为自主完成任务和实现人-机器人智能交互的前提.在规模较大的室内空间中,由于机器人常用的 RGB-D(RGB depth)视觉传感器(同时获取彩色图像和深度信息)视野有限,使之难以直接获取包含整个区域的单帧图像,但机器人能够运动到不同位置,采集多种视角的图像数据,这些数据总体上能够覆盖整个场景.在此背景下,提出了基于多视角 RGB-D 图像帧信息融合的室内场景理解算法,在单帧 RGB-D 图像上进行物体检测和物体关系提取,在多帧 RGB-D 图像上进行物体实例检测,同时构建对应整个场景的物体关系拓扑图模型.通过对 RGB-D 图像帧进行划分,提取图像单元的颜色直方图特征,并提出基于最长公共子序列的跨帧物体实例检测方法,确定多帧图像之间的物体对应关联,解决了 RGB-D 摄像机视角变化影响图像帧融合的问题.最后,在 NYUv2(NYU depth dataset v2)数据集上验证了本文算法的有效性.

关键词 物体检测;物体实例检测;RGB-D 图像;物体关系拓扑图;场景理解

中图法分类号 TP391.41

近年来,场景理解成为机器视觉以及智能机器人领域备受关注的一个重要问题.对于工作在人居环境的服务机器人来说,通过检测、识别场景中的各种物体并提取物体之间的关系,有助于更好地理解其所在环境,从而实现任务规划、物体搜索、人-机器人交互等自主行为.

可想而知,机器人必须全面地掌握环境中的信息,才能正确地理解环境.但是机器人搭载的视觉传感器,例如 RGB-D(RGB depth)传感器,通常视野有限,无法直接采集到覆盖整个场景的图像数据帧.因此,机器人需要游走到不同视角位置采集多帧数据,并将多帧数据中的信息进行综合,才能获得其所在场景的全面信息.

为实现多帧图像数据间的信息融合,常见的做法是通过图像拼接方法将多帧图像数据整合成为一幅全景图像[1],进而进行物体检测等信息提取工作.但是图像拼接算法[2]对视角变化较为敏感,效果不稳定.另外一些做法是通过 3 维重建算法[3-4]进行多帧数据的信息融合,这需要进行大量复杂计算,而服务机器人工作在动态变化多发的室内环境中,维护 3 维模型的计算及存储代价较高.

基于以上背景,本文提出了一种基于多视角 RGB-D 图像帧信息融合的场景理解算法.该算法的目标在于通过物体实例检测将不同图像帧的物体检测结果、关系提取结果进行融合,最终得到表征整个场景理解结果的物体关系拓扑图.该算法流程如图 1 所示,首先对多帧图像逐帧进行物体检测与关系提取,随后将每帧图像物体检测结果与场景中已知的同类物体进行实例检测,将重复的物体实例去除,将新发现的物体实例与原有的物体关系图进行融合,根据融合结果对物体关系图进行更新,以此迭代完成整个场景的物体检测和关系提取.

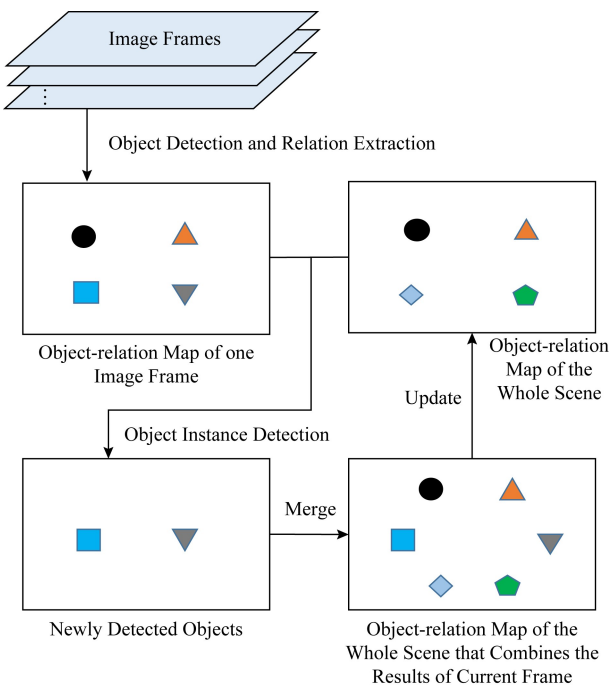


Fig. 1 The flow diagram of our algorithm

图 1 本文算法的流程图

本文主要贡献有 2 个方面:1)对 RGB-D 图像帧进行划分并提取图像单元的颜色直方图特征,基于此提出了基于最长公共子序列的跨帧物体实例检测算法,以支持多视角图像帧的融合;2)完成了面向室内

整体场景、融合多视角 RGB-D 图像帧的场景理解，并生成物体关系图作为场景理解的结果模型.

1 相关工作

1.1 物体检测

物体检测是计算机视觉领域的研究热点，也是机器人领域场景理解任务的关键问题.近年来，深度卷积神经网络(convolutional neural network，CNN)已经成为物体检测任务的基础工具[5].深度神经网络在提取图像高层抽象特征方面性能优越，早期的工作如 R-CNN[6](region-based convolutional neural network)以及其加速版本 Fast R-CNN[7]使用深度卷积神经网络提取特征，结合候选框提取算法实现物体检测.后续工作 Faster R-CNN[8-9]提出候选框提取网络(region proposal network，RPN)将候选框提取过程囊括至端到端的网络模型中.Mask R-CNN[10]在 Faster R-CNN 的基础上添加了用于分割物体实例的网络分支，实现了集物体检测和语义分割于一体的网络模型.

1.2 物体实例检测

物体实例检测算法的目标是检测出现在不同图像帧中的同一物体的实例.一些使用纹理特征(例如灰度共生矩阵[11])的算法通过描述物体表面纹理分布来进行物体实例判别.Bao 等人[12]使用基于部件的表示方式对物体实例进行建模，并通过物体的部件外观和几何特征进行比较来判断物体对象的外观一致性.通过特征点匹配进行物体实例检测[13-14]是另一种算法思路，这种算法对于平面的形变具有一定的鲁棒性，但同时受视角变化影响较大.

图像检索问题[15]与物体实例检测具有关联性，同时也存在明显差异.图像检索算法通常对图像整体进行相似度评估，而物体实例在不同视角下往往会在整体上产生较大的差异.

1.3 环境建图

地图构建是智能机器人领域的一个经典问题，近年以来，集成了空间几何信息和物体属性信息的语义地图成为主流的地图格式.Silberman 等人[16]将室内 RGB-D 数据解析为地板、墙壁、支撑面和物体等区域，然后恢复不同区域之间的支撑关系.Koppula 等人[17]利用各种特征和上下文关系，包括局部视觉外观、形状、物体共现和几何关系等，对点云数据进行语义标注，同时抽取不同语义点云之间的邻近或者上下关系.

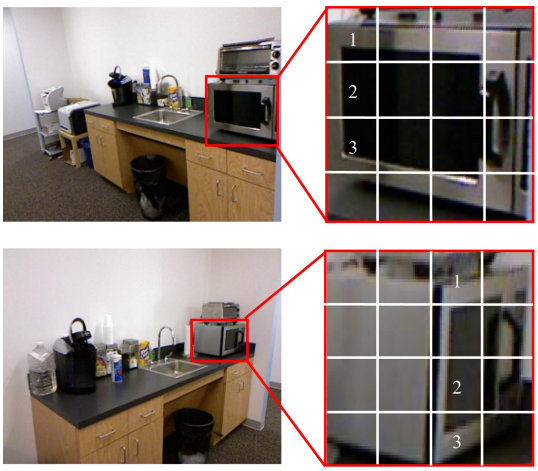
语义地图可用于导航和各种操作任务，但其存在存储消耗大、构建及维护计算量大的问题，且在环境发生动态变化时鲁棒性较差.因此 Ikehate 等人[18]定义了一系列的图结构生成室内环境的结构化模型，如以图节点表示房间、墙和物体等元素，以实线表示 2 个节点具有公共边界，以虚线表示没有边界约束的附着关系等，与语义地图相比，这种地图具有结构简单、易于构建和维护的优点.类似地，本文提出结构简单、易于构建和维护的物体-关系拓扑图作为场景理解的表达.

2 物体实例检测算法

本文算法在融合多图像帧中的物体检测结果时，需要识别出现在不同图像帧中的相同物体实例，物体实例是指在不同视角图像帧中出现的同一个物体的样本.在识别物体实例时，物体的颜色特征比结构特征更加具有区分性，但在不同图像帧中，物体的完整性与视角均可能发生变化，因此，同一个物体实例在不同图像帧中的整体颜色特征可能产生较大差异，在此背景下本文采用基于图像单元的颜色特征提出了物体实例检测算法.

2.1 图像划分和 LCS 算法描述

在不同视角下，物体实例的完整性与位姿均会发生变化，因此，同一物体实例的颜色直方图在不同图像中也会有较大差异.但同时，不同图像帧中的相同物体实例会具有一定的共同部分，为了更好地表达这种跨帧、跨视角的部分对应关系，划分为若干个细胞单元，则其细胞单元之间具有相应的对应关系，如图 2 所示：



Number 1, 2, 3 represent corresponding cells.

Fig. 2 The correspondence of some cells

图 2 部分单元对应关系

本文引入最长公共子序列算法(longest common subsequence, LCS)描述 2 个图像的细胞单元序列的相似程度[19-20].LCS 算法在文本分析中常用来计算 2 个字符串的相似度.对于 2 个字符序列$(S1,S2)$,记 $S1_i,S2_j$ 分别为序列 $S1,S2$ 的子序列$(S1_0,S1_1,\cdots,S1_i)$,$(S2_0,S2_1,\cdots,S2_j)$,则 $V_{\mathrm{LCS}}(S1_i,S2_j)$计算方式为

$$V_{\mathrm{LCS}}(S1_i,S2_j)=\begin{cases}V_{\mathrm{LCS}}(S1_{i-1}+S2_{j-1})+1, & S1_i=S2_j;\\ \max(V_{\mathrm{LCS}}(S1_i,S2_{j-1}),V_{\mathrm{LCS}}(S1_{i-1},S2_j)), & S1_i\neq S2_j.\end{cases} \tag{1}$$

LCS 算法可以通过动态规划进行优化,即引入状态数组对中间状态进行记录以避免多层递归.假设字符序列$(S1,S2)$,其长度分别为$(l1,l2)$,其状态数组记为 $c[l1,l2]$.在计算过程中,首先将数组的第 1 行和第 1 列初始化为 0,表示 2 个序列其中之一长度为 0 的情况.随后从 $c[1,1]$开始,按照从左向右、从上到下的顺序,计算数组中每个位置的值:

$$c[i,j]=\begin{cases}c[i-1,j-1]+1, & S1_i=S2_j;\\ \max(c[i-1,j],c[i,j-1]), & S1_i\neq S2_j.\end{cases} \tag{2}$$

最后,数组右下角的数值 $c[l1,l2]$即为序列$(S1,S2)$的最长公共子序列长度.在本文中,2 幅图像分别被均等划分为 n 个细胞单元,则在获取 2 个图像单元序列的最长公共子序列长度后,即可计算 2 个图像单元序列的相似度:

$$similarity(S1,S2)=V_{\mathrm{LCS}}(S1,S2)/n. \tag{3}$$

当 2 个图像子单元序列之间的相似度大于某阈值时,则认为 2 个图像序列代表的是同一个物体,实验中阈值选择为 0.55.

2.2 图像细胞单元相似度定义

图像单元并不像字符一样可以直接比较是否相等,因此本文提取图像细胞单元的颜色特征以描述 2 个图像单元的相似性.颜色直方图是反映颜色分布的重要特征.本文使用 RGB 颜色空间统计颜色直方图,以 R(红色)、G(绿色)、B(蓝色)这 3 种基本色为基础,进行不同程度的叠加来表示不同的颜色模型,其中每种基本色按亮度的不同分为 256 个等级.

由于在不同视角处光照条件也存在差异,因此同一物体实例在不同图像中成像亮度可能发生变化.为减少亮度变化带来的影响,在统计颜色直方图时将 256 个亮度等级平均划分为 8 个区间,每个区间大小为 32,即每个通道将产生一个 8 维直方图向量.随后,每个通道的直方图向量被归一化至区间$[0,1]$,记为 $\boldsymbol{r},\boldsymbol{g},\boldsymbol{b}$.为了更加精细地表达颜色分布,将 3 个直方图向量 $\boldsymbol{r},\boldsymbol{g},\boldsymbol{b}$ 两两对比,计算其差的绝对值,形成通道之间的颜色对比向量,计算方法为

$$cmp(\boldsymbol{\alpha},\boldsymbol{\beta})=\mathrm{abs}(\boldsymbol{\alpha}-\boldsymbol{\beta}). \tag{4}$$

将 3 个颜色通道 RGB 各自像素值的总和记为 $S_{\mathrm{R}},S_{\mathrm{G}},S_{\mathrm{B}}$,记 $S_{\mathrm{A}}=S_{\mathrm{R}}+S_{\mathrm{G}}+S_{\mathrm{B}}$,则图像的颜色直方图特征为

$$feature=(\boldsymbol{r},\boldsymbol{g},\boldsymbol{b},cmp(\boldsymbol{r},\boldsymbol{g}),cmp(\boldsymbol{r},\boldsymbol{b}),cmp(\boldsymbol{g},\boldsymbol{b}),S_{\mathrm{R}}/S_{\mathrm{A}},S_{\mathrm{G}}/S_{\mathrm{A}},S_{\mathrm{B}}/S_{\mathrm{A}}), \tag{5}$$

其中包含 6 个 8 维向量和 3 个实数.当 2 个图像单元进行比较时,对于颜色直方图特征中的 6 个 8 维向量,分别计算对应向量之间的欧氏距离,对于特征中的 3 个实数,分别计算对应实数之间差的绝对值,最终将计算结果串联为 9 维向量,该向量即用来描述 2 个图像单元间的相似性.通过训练支持向量机(support vector machine, SVM)来判断 2 个图像单元是否能够视为同一个“字符”.

3 融合多视角图像帧的场景理解

本文基于第 2 节给出的物体实例检测算法,实现基于多视角图像帧的场景理解,即通过物体实例检测算法将多帧图像中的物体检测结果进行融合,构建场景中的物体关系拓扑图.下面对物体检测和物体关系图构建 2 个方面的内容进行详细介绍.

3.1 基于 Mask R-CNN 模型的单帧图像物体检测

随着深度学习在物体检测领域的应用,各种深度网络模型被相继提出,物体检测的精度也逐渐提高.Mask R-CNN 是用于物体检测与语义分割的深度网络模型,该模型在 Faster R-CNN 的基础上添加了物体实例分割分值,因此能够同时进行物体检测与语义分割.本文使用在数据集上预训练好的 Mask R-CNN 模型,在 NYUv2(NYU depth dataset v2)数据集[15]上微调,进行物体检测.

NYUv2 数据集包含 1 449 张标注的 RGB-D 图像,来自 3 个城市中 484 个不同的商业和住宅区域.本文将该数据集手工划分为 610 个场景片段,表 1 展示了含有不同数量图像帧的场景片段分布情况.其中,包含图像帧数量为 1 的场景片段即为独立的单帧图像,包含图像帧数量超过 4 的场景片段多采集于商场等大型场景.考虑到本文的研究背景,将包含图像帧数量为 2,3,4 的图像划分为测试集,其他图像划分为训练集.本文在划分后的数据集上使用经过微调的 Mask R-CNN 模型进行物体检测.

Table 1 Statistics of Scene Fragments

表 1 分布情况

Number of Image Frames Included in the Scene	Number of Scenes
1	248
2	118
3	130
4	69
5	30
6	9
7	2
8	1
14	1
27	1
33	1

3.2 基于物体实例检测的多帧图像融合

对于同一场景采集的多帧图像,首先使用 3.1 节给出的算法对单帧图像进行物体检测,然后使用前述物体实例检测算法将单帧图像中的每个物体与当前场景中已有的物体进行比对以进行去重复处理,从而获取对场景中所有物体的检测与识别结果.

3.3 物体关系图构建

在基于多视角图像帧完成对整个场景中的物体进行检测识别之后,提取物体之间的关系,构成物体-关系拓扑图,作为场景理解结果的呈现形式.简单起见,本文仅考虑物体之间的相对位置关系.

随着深度传感器的普及使用,能够方便地获取场景的深度数据.深度图像包含了显式的空间信息,为提取物体之间的相对位置关系提供了方便.与文献[21]类似,本文定义了物体之间最常见的 2 类空间关系,分别是"邻近"与"上下"关系.在获取深度图像后,首先使用中值滤波对图像进行降噪以去除边界处的毛边,随后计算每个物体框之间的豪斯多夫距离.对于点集 A 和 B,豪斯多夫距离计算过程是:首先计算点集 A 中的任一点 a_i 到点集 B 中任一点 b_i 的最短距离 d_i,然后选取其中的最大值,即为点集 A 和 B 之间的豪斯多夫距离,表示为

$$H(A,B)=\max(a\in A)\{\min(b\in B)d(a,b)\}. \quad (6)$$

对于某个物体 O 而言,周围物体与其关系存在如图 3 所示的 2 种情况.

对于"邻近"关系,2 个相互邻近的物体在图像中可能表现为左右相邻和前后相邻,由于存在视角变化,左右相邻和前后相邻能够在不同视角下互相转化,因此统一表现为"邻近"关系.而"上下"关系不

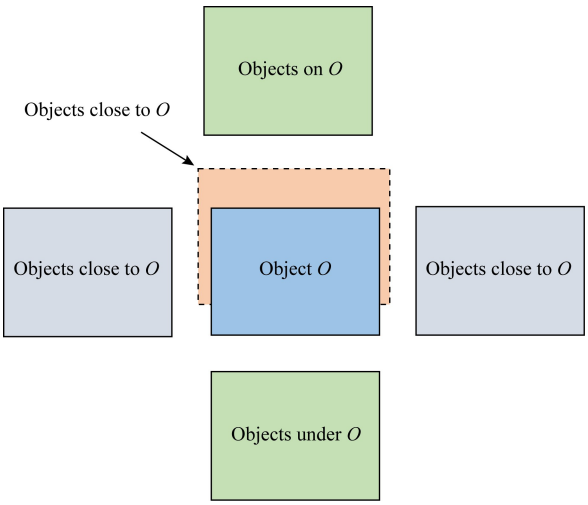


Fig. 3 The relationship between objects

图 3 物体关系示意图

受视角的影响,只存在一种情况.然而,对于"邻近"关系和"上下"关系,其整体豪斯多夫距离可能是十分相近的,因此对于 2 个物体框只计算水平方向和深度方向的距离,最后以 2 个物体框垂直方向上的中点距离以及垂直投影重叠度来判断其上下关系.同时,由于存在透视现象,距离传感器越远的物体在图像上成像越小,因此在计算水平方向的距离时设置了一个与深度大小相关的系数,即:

$$d(a(x_i,y_i,d_i),b(x_j,y_j,d_j))=$$
$$\|w\times\min(d_i,d_j)\times(x_i-x_j),(d_i-d_j)\|. \quad (7)$$

当 2 个物体框之间的豪斯多夫距离小于某个阈值 T_1 时,则认为其存在关系,否则认为其距离很远,相互之间影响力较小,不产生关系.当 2 个物体框靠近且 2 个物体框之间的垂直中点距离大于某阈值 T_2,并且二者水平向下的投影重叠度超过某阈值 T_3 时,判断其为"上下"关系,否则为"邻近"关系.

本文使用有向图结构构建物体关系图.对于物体关系,本文设置了语法规则构建物体关系图:

1) 有向图中每个节点表示一个物体;

2) 节点具有颜色属性,该属性表示了物体的类别信息;

3) 具有"邻近"关系的物体之间使用虚线无向边进行连接;

4) 具有"上下"关系的物体之间使用实线有向边进行连接,有向边由上方的物体指向下方的物体;

5) 不存在任何关系的物体之间不存在边.

可见,与经典的机器人地图相比,物体关系图不保留原始图像数据、不进行 3 维环境模型重建,因此所需计算量、存储量大大减少,且图的表达形式更为

直观,可以较好地支持语义导航、物品搜索等机器人典型任务.例如物品关系图中的邻近关系,可以帮助机器人根据已经发现的物品定位未知相关物体的搜索区域.

4 实验结果与分析

4.1 物体检测实验

本文使用在 NYUv2 数据集上微调的 Mask R-CNN 模型检测其中的物体.该实验中,物体类别共有 21 类,训练图像 832 张、测试图像 617 张.表 2 展示了物体检测实验效果:

Table 2 Results of Object Detection

表 2 物体检测结果展示

Classes	Precision	Recall	Number of Labeled Classes
Mantel	0	0	14
Counter	0.292 3	0.180 9	105
Toilet	0.617 6	0.617 6	34
Sink	0.491 2	0.595 7	47
Bathtub	0.250 0	0.105 2	19
Bed	0.661 2	0.574 7	214
Headboard	0.245 9	0.277 7	54
Table	0.409 2	0.274 0	354
Shelf	0.558 1	0.123 7	194
Cabinet	0.296 8	0.369 3	463
Sofa	0.435 3	0.615 8	164
Chair	0.370 8	0.667 8	557
Chest	0	0	181
Refrigerator	0	0	29
Oven	0.419 3	0.590 9	22
Microwave	0.523 8	0.523 8	21
Blinds	0.379 4	0.489 0	182
Curtain	0.383 5	0.269 2	104
Board	0	0	65
Monitor	0.284 1	0.584 2	89
Printer	0.160 0	0.173 9	23
Total	0.380 7	0.398 6	2 935

4.2 物体实例检测实验

本节描述了 2 个实验.第 1 个实验在 NYUv2 数据集中选取了 207 对正样本、142 对负样本,对比本文提出的基于图像细胞单元的颜色直方图特征算法与整体颜色直方图特征算法(记为 SimColor)和基于 ORB(oriented FAST and rotated BRIEF)特征匹配的算法(记为 ORB-Match)的实例检测效果,展示本文算法的有效性.正样本为出现在不同图像帧中同一物体实例对,负样本则为随机选取的不同物体实例.

整体颜色直方图算法统计样本对的整体直方图,并计算对比向量,然后训练 SVM 分类器来判断其是否为同一物体实例.基于 ORB 特征匹配[14]的算法通过统计样本对中高质量 ORB 特征匹配点的数量,判断是否同一物体实例,本文设定匹配点数量大于等于 4 即表示同一物体.

表 3 展示了不同算法的检测结果,由表 3 可见本文提出的算法在该样本集上表现出了更好的效果.此外,实验表明该算法对视角、光照条件等因素的变化所带来的影响具有更好的鲁棒性.

Table 3 Results on Positive-Negative Samples

表 3 正负样本集实验

Algorithm	Precision	Recall
SimColor	0.54	0.47
ORB-Match	0.67	0.51
Ours	0.94	0.95

第 2 个实验对 NYUv2 数据集中的 58 个场景片段进行了标注与实验,每个场景片段包含几帧不同视角的图像,包含 379 对物体实例对.表 4 展示不同算法检测结果的精确率与召回率,并使用 F-Score 对精确率与召回率进行综合评估.F-Score 是精确率(precision, P)和召回率(recall, R)的加权调和平均,常用来评估分类模型的好坏,计算为

$$F_\delta=\frac{(\delta^2+1)\times P\times R}{\delta^2\times P+R}, \tag{8}$$

其中 $\delta=1$.

Table 4 Results on NYUv2

表 4 综合实验

Algorithm	Precision	Recall	F-Score
SimColor	0.13	0.35	0.189 5
ORB-Match	0.37	0.52	0.432 3
Ours	0.58	0.72	0.642 4

实验结果如表 4 所示,可以得出,本文基于图像细胞单元的颜色直方图物体实例检测算法精确率和召回率均优于整体颜色直方图算法及基于 ORB 特征的匹配算法.

4.3 场景理解与关系图构建实验

本实验展示了实例检测结果和建立物体关系图

的过程.表 5,6 展示了融合计算过程,其中 Monitor1_added 代表已经加入到场景中的显示器实例 1,使用 Monitor1_f2 代表第 2 帧图像中的显示器实例 1.表 5,6 中仅列出对比的物体实例.表 7 给出了某个具有 3 帧图像的室内场景的理解与建图过程.

在表 7 中,第 1 帧图像有 4 个物体被检测出来.由于物体关系拓扑图初始状态为空,因此 4 个节点均被添加至关系图中,使用 3.3 节中描述的方法判断显示器 2 和打印机 1 是相邻的关系,显示器 2 和打印机 1 之间用无向虚线连接.

在第 2 帧图像中 4 个物体被检测到,其中显示器和打印机已经出现,因此使用第 2 节描述的物体实例检测算法比较同类别物体的相似性.从表 5 中可看到,第 2 帧中的显示器与场景中已有的显示器实例 1 匹配度更高,且大于阈值 0.55,所以认为第 2 帧图像中的显示器为已有显示器实例 1,打印机实例的判别过程类似.使用 3.3 节中描述的方法判断打印机 1 位于桌子 1 上,所以打印机 1 和桌子 1 之间使用有向实线连接,箭头由打印机指向桌子,其余判定相邻的物体使用虚线连接.第 3 帧图像的融合过程类似,融合计算参考表 6.

Table 5 Multi-Frame Fusion Process Step 1

表 5 多帧融合过程-第 1 步

Object in Scene	Similarity	
	Monitor1_f2	Printer1_f2
Monitor1_added	0.57	
Monitor2_added	0.41	
Printer1_added		0.59

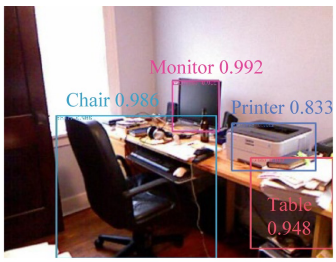
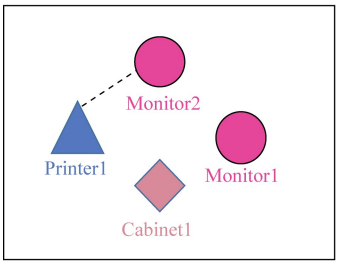
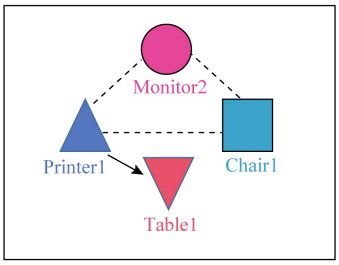
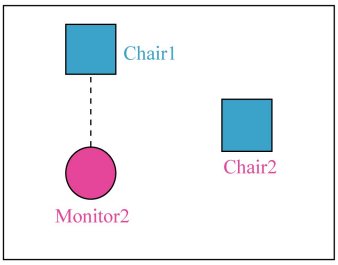
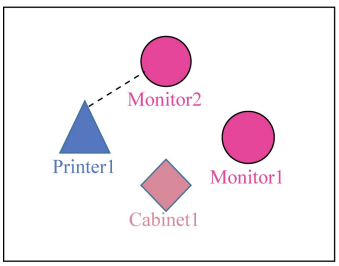
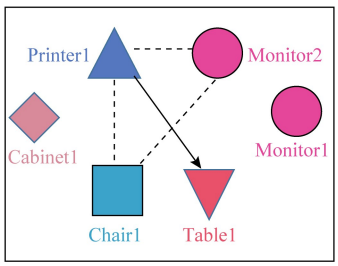
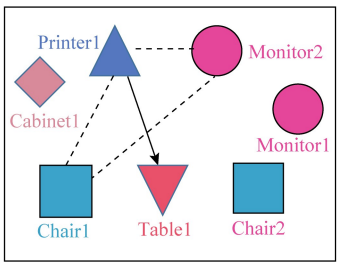
Table 6 Multi-Frame Fusion Process Step 2

表 6 多帧融合过程-第 2 步

Object in Scene	Similarity		
	Monitor1_f3	Chair1_f3	Chair2_f3
Monitor1_added	0.81		
Monitor2_added	0.38		
Chair1_added		0.77	0.40

Table 7 Result of Scene Understanding and Map Building

表 7 场景理解与建图结果

Phase	The First Frame	The Second Frame	The Third Frame
Object Detection			
Object-relation Map of Single Frames			
Object-relation Map of Scene			

Notes: "- - - - - -" represents the close-to relation and "——→" represents on-top-of relation from object A to object B if object A is on the top of object B.

5 总 结

本文提出了一种直观有效的跨帧物体实例检测算法,并在此基础上实现了基于多视角 RGB-D 图像帧信息融合的室内场景理解,融合多帧图像中的物体检测结果,并提取物体间的拓扑关系,构建场景内的物体关系图.该算法能够有效集成多帧图像中的场景信息,实现整体场景理解.如何进一步提高算法的实时性,并应用于移动机器人的在线运行及具体任务,是下一步的研究方向.

参 考 文 献

[1] Wang Maosen, Niu Shaozhang, Yang Xuan. A novel panoramic image stitching algorithm based on ORB [C] // Proc of the 2017 IEEE Int Conf on Applied System Innovation. Piscataway, NJ: IEEE, 2017: 818-821

[2] Ghosh D, Kaabouch N. A survey on image mosaicing techniques [J]. Journal of Visual Communication and Image Representation, 2016, 34: 1-11

[3] Dang T K, Worring M, Bui T D. A semi-interactive panorama based 3D reconstruction framework for indoor scenes [J]. Computer Vision & Image Understanding, 2011, 115(11): 1516-1524

[4] Chen Kang, Lai Yukun, Hu Shimin. 3D indoor scene modeling from RGB-D data: A survey [J]. Computational Visual Media, 2015, 1(4): 267-278

[5] Liu Li, Ouyang Wanli, Wang Xiaogang, et al. Deep learning for generic object detection: A survey [J]. arXiv preprint, arXiv:1809.02165, 2018

[6] Girshick R, Donahue J, Darrelland T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] //Proc of the 2014 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 580-587

[7] Girshick R. Fast R-CNN [C] //Proc of the 2015 IEEE Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1440-1448

[8] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149

[9] Huang Jipeng, Shi Yinghuan, Gao Yang. Multi-scale Faster-RCNN algorithm for small object detection [J]. Journal of Computer Research and Development, 2019, 56(2): 319-327 (in Chinese)
(黄继鹏, 史颖欢, 高阳. 面向小目标的多尺度 Faster-RCNN 检测算法[J]. 计算机研究与发展, 2019, 56(2): 319-327)

[10] He Kaiming, Gkioxari G, Dollar P, et al. Mask R-CNN [C] //Proc of the 2017 IEEE Int Conf on Computer Vision, Piscataway. NJ: IEEE, 2017: 2980-2988

[11] Rui Yong, Huang T, Ortega M, et al. Relevance feedback: A power tool for interactive content-based image retrieval [J]. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 644-655

[12] Bao S Y, Xiang Yu, Savarese S. Object co-detection [C] // Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2012: 86-101

[13] Lowe D G. Object recognition from local scale-invariant features [C] //Proc of the 7th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 1999: 1150-1157

[14] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF [C] //Proc of the 2011 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2011: 2564-2571

[15] Zhou Wengang, Li Houqiang, Tian Qi. Recent advance in content based image retrieval: A literature survey [J]. arXiv preprint, arXiv:1706.06064, 2017

[16] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images [C] //Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2012: 746-760

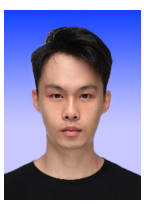
[17] Koppula H S, Anand A, Joachims T, et al. Semantic labeling of 3D point clouds for indoor scenes [C] //Proc of the 25th Annual Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2011: 244-252

[18] Ikehate S, Yan Hang, Furukawa Y. Structured indoor modeling [C] //Proc of the 2015 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1323-1331

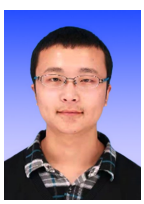
[19] Ullman J D, Aho A V, Hirschberg D S. Bounds on the complexity of the longest common subsequence problem [J]. Journal of the ACM, 1976, 23(1): 1-12

[20] Nakatsu N, Kambayashi Y, Yajima S. A longest common subsequence algorithm suitable for similar text strings [J]. Acta Informatica, 1982, 18(2): 171-179

[21] Lin Dahua, Fidler S, Urtasun R. Holistic scene understanding for 3D object detection with RGBD cameras [C] //Proc of the 2013 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2013: 1417-1424

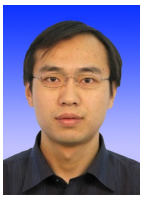


Li Xiangpan, born in 1997. Master. His main research interests include computer vision, scene understanding.



Zhang Biao, born in 1996. Master candidate. His main research interests include computer vision, scene understanding.

Sun Fengchi, born in 1973. PhD, associate professor. Member of IEEE. His main research interests include artificial intelligence, robotics and embedded system.



Liu Jie, born in 1979. PhD, professor. Member of CCF. His main research interests include machine learning, data mining and natural language processing.

2020 年《计算机研究与发展》专题(正刊)征文通知
——人机混合增强智能的典型应用

当前,以大数据驱动的机器学习为核心的人工智能在不确定性、脆弱性和开放性实际应用环境中面临重大挑战.随着知识引导的兴起,一种新的学习范式——“知识引导+数据驱动”的人机混合增强智能应运而生.其基本思路是将人的高级认知、推理和随机决策能力引入到机器高效的计算过程中,实现人在回路的混合增强智能.人机混合的增强智能还面临一系列难题,例如,如何将人的认知、决策行为与机器的知识表征、因果推理过程有效融合;如何构建面向不同计算应用任务的人机混合智能增强智能方法;如何表征和评估人与机器形成的混合增强智能系统的性能等.

《计算机研究与发展》拟于 2020 年 12 月出版应用技术专题——人机混合增强智能的典型应用.本专题希望围绕上述难题讨论人机混合增强智能的关键技术与发展趋势,报导相关技术在行业中的实践案例,交流思想和成果,进而促进相关技术的研究与发展.

征文内容 本专题包括(但不限于)下列主题:

1) 人机混合的知识表征与融合;
2) 人机混合的知识理解与因果推理;
3) 人机混合增强智能在教育领域的典型应用;
4) 人机混合增强智能在舆情分析领域的典型应用;
5) 人机混合增强智能在智慧税务领域的典型应用;
6) 人机混合增强智能在智慧医疗领域的典型应用.

投稿要求

1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表或宣读过,不存在一稿多投问题.作者在投稿时,需向编辑部提交版权转让与投稿声明.
2) 论文一律用 Word 排版,格式体例请参考《计算机研究与发展》近期文章.
3) 论文请通过期刊网站 (http://crad.ict.ac.cn)进行投稿,并在作者留言中注明“人机混合增强智能 2020 专题”(否则按自由来稿处理).

重要日期

征文截止日期: 2020 年 9 月 15 日
录用通知日期: 2020 年 10 月 15 日
修改稿提交日期: 2020 年 10 月 20 日
出版日期: 2020 年 12 月

特邀编委

郑庆华 教授 西安交通大学 qhzheng@mail.xjtu.edu.cn

联系方式

编 辑 部:crad@ict.ac.cn, 010-62620696, 010-62600350
通信地址: 北京 2704 信箱《计算机研究与发展》编辑部
邮 编:100190