RGB-D 图像分类方法研究综述

涂淑琴[1] 薛月菊[2] 梁 云[1] 黄 宁[2] 张 晓[2]

[1]华南农业大学数学与信息学院，广东 广州 510642

[2]华南农业大学电子工程学院，广东 广州 510642

摘要 采用新型 3D 传感器能够便捷地同时获取多场景、多视觉和多目标彩色和深度信息的 RGB-D 图像，利用其在物体重叠和遮挡下深度信息对颜色和亮度的不变特点，有效提高 RGB-D 图像分类的精度。对微软 Kinect 设备的发展及原理做详细介绍；介绍了现有的 RGB-D 数据集；对现有 RGB-D 图像特征提取与分类方法进行了归纳、分析和比较；阐述 RGB-D 图像分类的发展趋势。

关键词 图像处理；目标识别；场景分类；特征提取；Kinect；RGB-D 图像

中图分类号 TP391 **文献标识码** A

doi：10.3788/LOP53.060003

Review on RGB-D Image Classification

Tu Shuqin[1] Xue Yueju[2] Liang Yun[1] Huang Ning[2] Zhang Xiao[2]

[1] *College of Mathematics and Informatics, South China Agricultural University, Guangzhou, Guangdong* 510642, *China*

[2] *College of Electronic Engineering, South China Agricultural University, Guangzhou, Guangdong* 510642, *China*

Abstract The color and depth information of multi-scenario, multi-vision and multiple target in the RGB-D images are conveniently obtained using a new 3D sensor at the same time. The RGB-D image classification accuracy is effectively improved using the depth information invariant characteristics of color and brightness, when the objects overlap and occlusion occurs. The development and theory of Microsoft Kinect are introduced in detail, and the existing RGB-D datasets are described. Then the feature extraction and classification methods are summarized, analyzed and compared. The development trend of RGB-D image classification is discussed.

Key words image processing; object recognition; scene classification; feature extraction; Kinect; RGB-D image

OCIS codes 100.3008; 100.6890; 100.5010; 130.6010; 150.6910

1 引 言

RGB 图像分类是计算机视觉中重要的基础问题，已广泛应用于国防和民用的许多领域。但在实际应用中，RGB 图像在目标重叠、遮挡、光照变化大、阴影和场景复杂等情况下，存在目标识别率低、场景分类效果不佳及稳健性差等问题。为克服这些困难，近几年利用 RGB-D 图像进行目标识别和场景分类的研究引起科技界、工程界和商业界极大兴趣[1-4]。2010 年以来，如 ASUS Xtion 和 Mesa SwissRanger 等新型 3D 传感器的出现，特别是微软推出价格适中的 Kinect 传感器[5]，极大地促进 RGB-D 图像在不同领域的研究应用。不随亮度、颜色变化而改变的深度信息，为复杂问题的场景及目标分类提供有用的额外信息，且目标几何信息使物体可以更好地从背景中被识别。研究表明，融合深度信息的 RGB-D 图像分类[6-22]具有更好的分类准确率及稳健性。

收稿日期：2016-01-04；**收到修改稿日期**：2016-02-25；**网络出版日期**：2016-06-01

基金项目：广东省科技计划(2015A020209148，2015A020224038，2015A020209124，2016A050502050)

作者简介：涂淑琴(1978—)，女，博士研究生，主要从事图像场景分类和目标识别方面的研究。

E-mail：tushuqin@163.com

导师简介：薛月菊(1969—)，女，教授，博士生导师，主要从事计算机视觉、智能计算方面的研究。

E-mail：xueyueju@163.com

RGB-D 图像分类是计算机视觉中的研究热点问题之一。2011 年至 2015 年计算机重要会议和期刊如 CVPR、ECCV、ICCV、NIPS、PAMI 和 IJCV 等，大量报道了关于 RGB-D 图像研究论文[23-36]。研究者们利用 RGB-D 图像进行了多种计算机视觉基本问题的理论和应用研究，包括 3D 重建[37-38]、对象识别与 RGB-D 图像分类[1-4,39-40]、室内 3D 地图构建[41-42]、身体跟踪[43-45]、手势分析[46]等。其中，RGB-D 图像分类主要是通过提取彩色图像和深度图像的特征建立分类模型，然后利用建立好的分类模型进行目标识别或者场景分类[1,9]。如何融合深度信息获取表达能力强的特征描述子是 RGB-D 图像分类性能优劣的重要因素，是复杂场景下获取稳健性强的 3D 分类识别系统的关键[11-20]。本文将在后文详细介绍和分析 RGB-D 图像特征提取及分类算法。

同时，研究者利用 3D 传感器(如 Kinect)建立了多场景、多种类目标及存在遮挡的 RGB-D 数据集[23-32]，包含日常生活中常见的目标物体，室内场景，并从多个视角拍摄，能够应用于大规模的 3D 目标识别和场景实际应用，促进 RGB-D 图像分类算法的快速发展。目前，除文献[47]外，关于 RGB-D 图像分类的综述文章极为少见，为帮助研究者对 RGB-D 图像分类研究工作有进一步了解，本文首先对目前流行的 Kinect 工作原理进行简要介绍，然后对现有的 3D 图像数据库进行详细说明，重点对 RGB-D 图像分类方法进行综述，最后对未来 RGB-D 特征融合及发展趋势进行展望。

2 Kinect 传感器基本原理

2.1 Kinect 基本功能

Kinect 传感器利用 Prime Sense 设备、飞行时间测量技术(TOF)技术和感应芯片，以低廉的价格实现深度摄像头功能，包括 Kinect 1.0 和 Kinect 2.0 两代产品，主要性能比较如表 1 所示。

Kinect 1.0 在 2010 年发布，结构如图 1 所示，包括彩色摄像头(640 pixel×480 pixel)，深度摄像头(320 pixel×240 pixel)，只能识别两个人的骨骼。Kinect 2.0 在 2014 年发布，结构如图 2 所示，包括一个深度传感器，一个红外(IR)发射器和 1080 p 彩色摄像机，4 阵列麦克风。与 Kinect 1.0 相比，Kinect 2.0 性能显著改进，能同时识别 6 个人的骨骼，新增活跃红外线，在黑暗及光线不足环境下能获取稳定的深度图，且识别精度高。

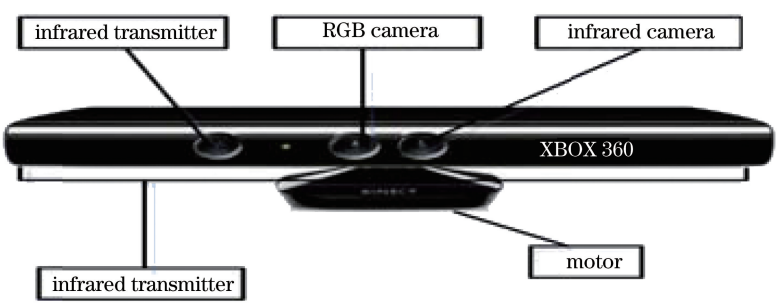

图 1 Kinect 1.0 结构图

Fig. 1 Structure of Kinect 1.0

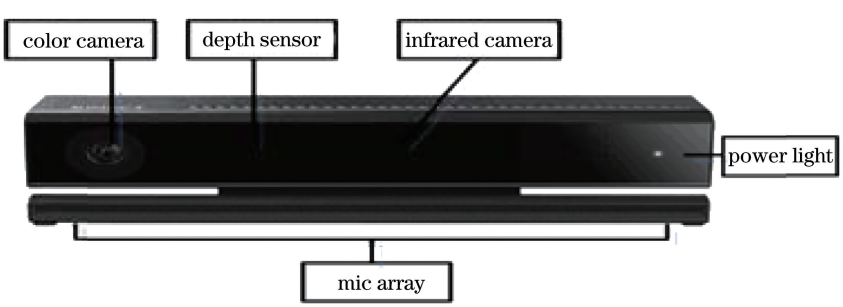

图 2 Kinect 2.0 结构图

Fig. 2 Structure of Kinect 2.0

表 1 Kinect 1.0 和 Kinect 2.0 的区别

Table 1 Difference between Kinect 1.0 and Kinect 2.0

Function	Kinect 1.0 (released in 2010)	Kinect 2.0 (released in 2014)
Field of view	Horizontal 57.5°, vertical 43.5°	Horizontal 70°, vertical 60°
Range	0.8～4.0 m	0.8～4.0 m
Video stream (color)	640 pixel×480 pixel×24 pixel 4∶3 RGB@30 frame/s; 640 pixel×480 pixel×16 pixel 4∶3 YUV@15 frame/s	1920 pixel×1080 pixel×16 pixel 16∶9 YUY2@30 frame/s
Video stream (depth)	320 pixel×240 pixel×16 pixel, 13 bit depth	512 pixel×424 pixel×16 pixel, 13 bit depth
Active infrared video stream	Null	512 pixel×424 pixel, 11 bit active range
Register	Colordepth	Colordepth and active infrared
Audio capture	4 array mic 48 Hz audio	4 array mic 48 kHz audio
Interface	USB 2.0	USB 3.0
Delay	Tape handling 90 ms	Tape handling 60 ms
Read depth	Light measurement	Flight time calculation
Tilt motor	Vertical only	Null

2.2 Kinect 获取深度信息的原理

Kinect 1.0 传感器利用 PrimeSense 公司的光编码(LC)技术来获取深度信息。光编码采用红外激光发射相对随机但又固定的斑点图案,将这些光斑打在物体上,因为与摄像头距离不同,所以被摄像头捕捉到的位置也不尽相同。利用一个标准的 CMOS 图像传感器,通过右侧的红外线接收器采集经过编码的红外光谱影像,最终将这些信息传输给 Prime Sense 的 PS1080 Soc(系统级芯片)。并对 Kinect 的原始参数进行一系列复杂的逻辑运算,将红外光编码的影像解码后就可以得到视野中的三维深度信息,从而生成场景的深度图像。其中,PS1080 芯片是 Kinect 1.0 的核心处理单元,控制红外光的发射、接收、编码、解码等过程,采用 USB 2.0 接口以一个同步时序传送深度信息。

深度图中每个像素的位深度为 11 bit,但不是所有的位都用来编码深度信息,当超出距离范围时,深度最大值被设置为 $V_{\max}=1024$,最小值为 $V_{\min}=290$,实质上只有 794 个深度值(10 bit)来编码每个像素的深度信息。

原始深度值 v 和距离 d[48]之间的关系为

$$d=\frac{8\times B\times F_x}{V_{\max}-v}, \tag{1}$$

式中 $B=0.075$ m,对应红外发射器和红外摄像机之间的距离(基线),F_x 是红外摄像机在水平方向的焦距长度,若 d 是负值,可以忽略。图 3 显示了 Kinect 深度数据特点,其中蓝色曲线是深度图中的 v 与 d 距离值间的关系,红色线表示传感器的最小测量深度,绿色区域是 Kinect 中建议的合理使用范围,黄色区域是研究者们拍摄物体使用的距离范围。获取的深度图分辨率会有很多噪声,因此需要使用去噪方法进行深度图预处理。同时,Kinect 1.0 计算斑点位移需要用图像在一个小范围区域内做块匹配,丢失了像素级别的细节。对凸凹不平的表面、物体边缘、很细的物体很难检测准确的深度信息,比如 Kinect 1.0 对水杯的把手等东西很难识别,人脸也很容易蜕化成一个球状物体,圆形的东西边缘不够圆滑。

Kinect 2.0 采用飞行时间测量技术摄像机,通过发射一个强度随时间周期变化的正弦信号,获得发射、接受信号的相位差及光脉冲之间的传输延迟时间来计算深度信息。具体过程为:由红外投影仪主动投射近红外光谱,使其照射到粗糙物体、或是穿透毛玻璃后,扭曲光谱,形成随机的反射斑点(即散斑);光谱被深度

摄像头读取后，深度摄像头分析读取到的红外光谱并生成深度图。由于 Kinect 2.0 中的 TOF 在一定程度上可以做到逐个像素的计算，同时能过滤背景噪声，因此获取的深度图像具有更高的分辨率和精度。

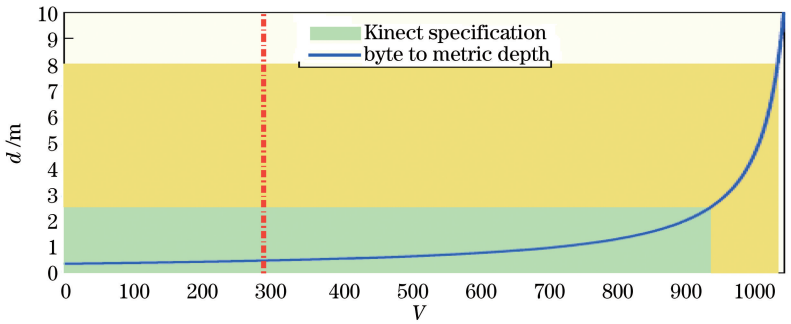

图 3 Kinect 深度数据

Fig. 3 Kinect depth data

3 RGB-D 数据库

为了进行 RGB- D 目标识别与场景分类研究，一些科研机构提供目标识别与场景分类的 RGB-D 测试数据集。目前主要使用的数据集包括：RGB-D object Dataset[23]、Berkeley 3D Object Dataset[24]、NYU Depth dataset[25,49]、Cornell RGBD dataset[26]、SUN3D[27]、SUN RGB-D[28]和其他的 3D 数据集[29-32,41-42]。

1) RGB-D dataset。该数据库由 Lai 等[23]利用 RGB-D 传感器对日常用品进行拍摄，是一个大规模、多层次和多视角的物体与场景数据库。将物体放置于受控的转台上，用 Kinect 传感器分别从与水平方向 30°、45°和 60°三种角度，同步拍摄其彩色和深度信息。数据集分两大部分：300 个家用物体构成的 51 个类别数据集及 8 个办公室和厨房视频 RGB-D 场景数据集。

2) Berkeley 3D Object Dataset。由 Janoch 等[24]发布，利用 Kinect 传感器在自然环境下拍摄，获取具有多视角和光照变化的家庭和办公环境下常见物品图像集。部分数据集如图 4(a)所示。

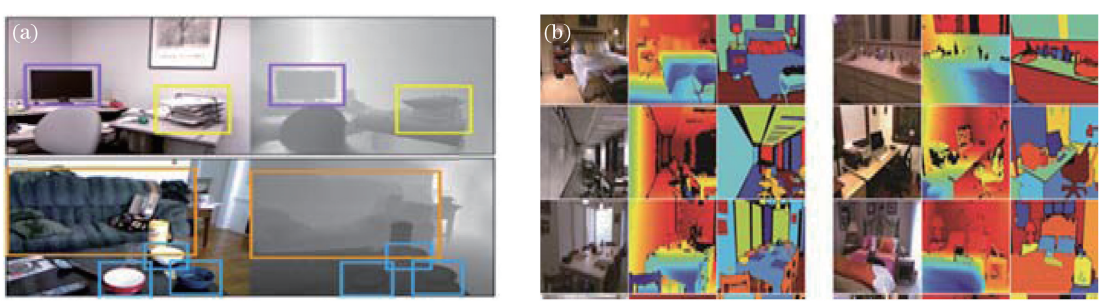

图 4 常见的部分 RGB-D 数据集。(a) B3DO 的部分实例；(b) NYU 深度数据集上部分场景数据

Fig. 4 Part of common RGB-D datasets. (a) Part of B3DO examples; (b) part of data in NYU depth dataset

3) NYU Depth dataset。由纽约大学 Silberman 和 Fergus 建立的一个极具挑战的室内场景分类数据库。包括 NYU Depth v1[49]和 NYU Depth v2[25]，其中 NYU Depth v1 包含 12 类 64 种不同室内场景，共 2347 幅图像；NYU Depth v2 包含 464 种不同场景共 1449 幅图像。该数据集有以下几个特点：① 通过微软 Kinect 传感器对各种场景进行拍摄，并采用特定的校正技术对深度图进行修正；② 用 Cross-bilateral 滤波器很好地滤除深度图中的阴影区域；③ 在 Kinect 摄像头上加上三轴加速度传感器，消除采集样本过程中出现的倾斜和晃动现象。因此，该数据集对于评估场景分类方案是非常合适的。部分数据集如图 4(b)所示。

4) Cornel RGBD。包含 52 室内场景数据，其中 24 个是办公室，28 个是家庭场景，每个场景采用 8～9 个 RGB-D 视角重构建成，并考虑物体之间的上下文关系，使用了 RGBDSLAM 将原始的 RGB-D 图像转化成三维点云，采用不同颜色进行物体标注。部分数据集如图 5 所示。

5) SUN3D。包含 415 段 RGB-D 视频流，在 41 幢不同大楼共 254 个不同面积地方进行拍摄，该视频流

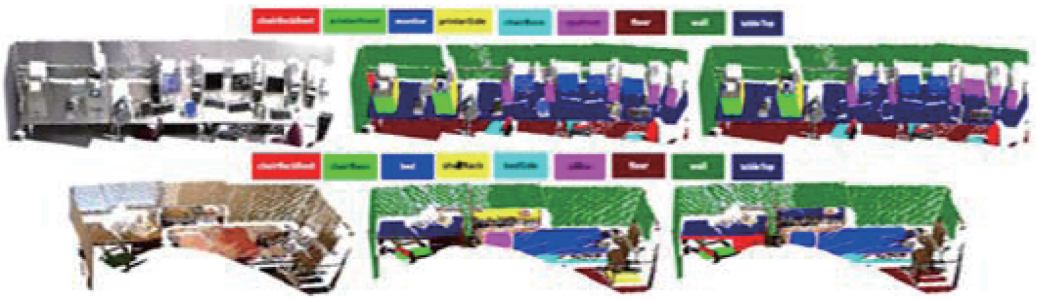

图 5 Cornel RGBD 部分室内场景

Fig. 5 Part of the indoor scene about Cornel RGBD

采用在某些关键帧上进行二维多边形标注，目前完成 8 个视频流的标注。主要优点是利用运动恢复结构(SFM)方法进行图像动态重构及多视觉深度修正方法改进深度信息。

6) SUN RGB-D。由 Song 等[28] 在 2015 年发布，采用 Intel Realsense、Xsus Xtion、Kinect v1 和 Kinect v2四种不同传感器采集数据，包含 47 类场景、800 个具体物体，共 10335 幅 RGB-D 图像，平均每幅图像中包含 14.2 个物体。对每幅图像中的物体都进行了人工标注，共包含 146617 个二维多边形标注、64595 个三维包围盒标注，同时对室内物体完成三维多边形空间布局。

7) 其他 RGB-D 数据集。还有一些数据集用于特殊的应用，如：人脸分析[29]，人姿态分析[30,33]，手势识别[32]，地图构建[41-42]，行为分析[35-36]等。

4 RGB-D 图像分类方法

RGB-D 图像分类处理流程如图 6 所示。其核心问题是如何利用彩色信息和深度信息进行目标特征描述，目标特征描述子优劣决定图像分类性能。根据特征描述子和特征学习方法不同，将 RGB-D 图像分类的

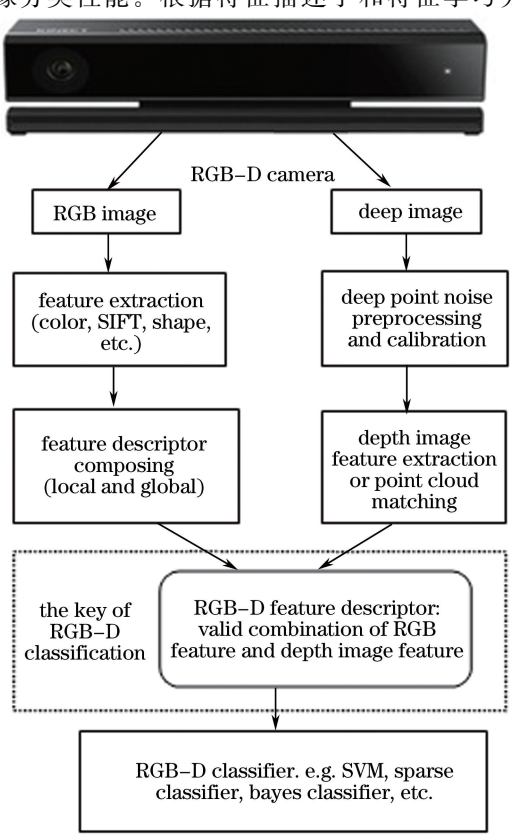

图 6 RGB-D 分类流程图

Fig. 6 Flow chart of RGB-D classification

方法分为4类。其中,二维手工设计特征和二维自动特征学习属于二维图像特征分类算法,全局三维特征描述子和局部三维特征描述子属于三维图像特征分类方法。这4类算法的共同之处为,利用深度信息提取对象深度图像特征,然后融合彩色图像特征描述子进行目标图像分类。表2列出了主要的特征提取方法和各方法的优缺点。

4.1 基于二维图像特征的RGB-D分类

基于二维图像特征RGB-D分类算法均为在原有二维图像识别算法基础上,对提取的彩色图像和深度图像特征进行分类识别。该类方法具有算法成熟的优点,已被广泛使用。主要包括:基于RGB-D的人工设计特征提取与表达方法及自动特征学习方法。

4.1.1 基于人工设计特征提取和表达的RGB-D分类

根据先验知识进行人工设计特征提取方法,包括尺度不变特征变换(SIFT)、加速稳健特性(SURF)和方向梯度直方图(HOG)等方法,然后采用词袋模型、空间金字塔匹配和稀疏编码方法进行特征表达。这些方法在原有RGB图像特征表达基础上,融合深度图像特征描述子,采用分类算法如SVM、贝叶斯分类器等进行图像识别。

表2 RGB-D图像特征提取与分类方法

Table 2 Image feature extraction and classification methods of RGB-D

RGB-D image classification method	RGB-D feature method	Feature description method	Characteristic
RGB-D classification based on 2D image feature	2D manual representation feature	Extracting depth and color image features with prior knowledge, and representing them using bag of words model, spatial pyramid matching and sparse representation	Simple algorithm and mature technologies. Extracting class features by manual design. Strict prior knowledge for specific visual problem is needed. Ignoring pose and position information of objects
	2D automatic feature learning	Automatically learning features using single layer CNN or multi layer networks in deep learning	General learning features automatically from source image. Computing expensive
RGB-D classification based on 3D image feature	Global 3D feature description	Obtaining appearance, shape and position information from segmented image by clustering point cloud	Applied in object recognition in complex scenes. Having higher accuracy and robust. Ability to recognize multi-objects simultaneously. Sensitive to scene blocking and noise
	Local 3D feature description	Extracting local 3D image features by pair-wise feature matching in real scene data	Good recognition in complex scenes. Ability to recognize objects under noise and changes of geometry. Expensive in recognition and pose restoring

Bo等[1]采用稀疏编码的特征提取方法,学习8个不同通道特征,包括灰度梯度,RGB特征,深度特征和形状特征,组成字典后,采用两层层次匹配追踪(HMP)进行特征表达,每层包含三个模块:正交匹配追踪,金

字塔和对比度归一化,获得较好的分类性能。其缺点是特征维度达到 188 和 300,导致分类运行时间长。Lai 等[2]提出实例距离学习(IDL)算法,能同时从多个特征角度选择最优距离特征,融合彩色图像特征和深度图像特征后,在 RGB-D 数据集上[23],其类别识别(判别一个目标属于哪个类)和实例识别(识别不同的目标实例)结果分别为 85.4%和 91.3%。Silberman 等[25]对 RGB 图像和对应的深度图像同时提取局部 SIFT 特征,利用空间金字塔匹配法 SPM 表达方法获取了最终用于表示样本的特征,该方法有效改善 RGB-D 场景分类性能。

Bo 等[3]在核函数学习特征基础上,提出 5 种深度核描述子,分别融合彩色与深度图像几个重要的特征,例如大小,形状和边缘信息。在 RGB-D 数据集上,其 RGB-D 类别识别和实例识别结果分别为 86.2%和 84.5%,对比 RGB 分类结果,类别识别和实例识别分别提高了 8.6%和 6%。BLum 等[4]提出了采用无监督方法学习 K 类均值,通过兴趣点监测 SURF 特征,从深度图像中提取一组卷积 K-均值描述符(CKM)特征,在 RGB-D 数据集上,其实例识别精度达到 90.4%,远远优于文献[3](实例识别:84.5%)的性能。

Ren 等[50]利用核描述提取局部梯度、颜色信息和深度梯度等特征信息后,采用超像素 MRF 和 gPb-ucm 分割方法进行场景分类,对比文献[25],其在 NYU RGB-D 数据库中分类准确率提高 20%。Cheng 等[51]提出半监督学习框架,采用 20%带类别标记训练集数据学习彩色和深度图特征,有效提高 RGB-D 目标识别率。

这些方法依靠手工设计提取初级与高级特征,对于具体的视觉问题需要具有很强的先验知识才能设计好区分性强的特征与融合规则,所以很难具有普适性。其次特征的参数需要大量的时间调节才能达到预期的效果,特征设计的好坏直接影响着整个系统的性能。因此,通用性强的特征提取方法引起研究者的极大兴趣,近年自动特征学习方法应运而生。

4.1.2 基于特征学习的 RGB-D 图像分类

近几年来,特征学习在 RGB 图像分类,目标检测和场景分类中得到广泛的应用。特征学习,也称表达学习[52],从原始图像出发,采用深度学习中不同的网络模型,通过无监督或有监督的训练,自动学习源图像中低层和高层特征。深度学习对多变化视觉问题具有普适性,能自动学习图像高层语义特征。常用特征学习结构包括:稀疏自动编码器结构、卷积神经网络(CNN)和多种深度网络结构融合。

Wang 等[53]利用二层无监督自动编码器结构学习特征,直接从 RGB 和深度图像学习和编码低层及高层特征,在 NYU N1 数据库上获得很好的场景分类性能。该方法优点是利用稀疏编码自动融合 RGB 和深度图特征,改变之前 RGB 和深度图各自独立提取特征的单模式特点。该模式能挖掘彩色和深度特征内在联系,有效地增强 RGB-D 图像分类的效果。

Socher 等[6]提出卷积和递归神经网络(CNN-RNN) 融合模型去学习 RGB-D 特征,原始的图像直接输入 CNN,获取低层特征,这些低层特征输入 RNN,通过抽象得到有效的高层特征,最后采用 softmax 方法分类,获得较好的分类性能。Schwarz 等[33]融合标准视图渲染深度信息,利用预训练颜色特征 CNN 模型,自动高效完成类别、实例和姿态分类。

Couprie 等[7]将 RGB 和深度图像作为 4 通道,并作为多尺度 CNN(由 Farabet 等[8]提出)的输入,在 NYU D2 场景数据进行场景分类,获得较好的性能; Gupta 等[22]研究基于 RGB-D 图像的轮廓检测、自底向上的分组问题以及语义分割问题。他们将用于普通图片分割的 gPb-ucm 方法扩展到了 RGB-D 图像,用于实现物体边缘的检测以及层次化的分割。他们在此基础上对物体重力方向进行预估计,采用卷积神经网络模型,融合物体深度图像水平视差、高度和多角度的梯度作为特征描述子,采用随机森林和 SVM 进行图像分类。对比直接采用 CNN 学习深度图像特征方法,该方法获得显著提升,其 RGB-D 图像识别准确率提高 56%。

Eitel 等[54]在深度图存在较多噪声和遮挡情况下,采用如图 7 所示的多模式 CNN 进行 RGB-D 识别,将彩色(蓝色)和深度图(绿色)分别输入 CNN 模型中,经过五层卷积和下采样操作,分别形成 256 个特征映射;然后采用 2 层全连接形成 4096 维特征,将彩色和深度特征经过一个全连接融合(灰色)层 fc1-fus 后,采用 softmax 分类器输出 51 个类别。

令 $D=\{(x^1,d^1,y^1),\cdots,(x^N,d^N,y^N)\}$ 为多模式 CNN 的输入;x^i,d^i,y^i 分别是 RGB 图像、深度图图像和图像类别标记; $g^I(x^i;\theta^I)$ 和 $g^D(d^i;\theta^D)$ 分别是 fc7 层彩色图和深度图的特征输出;W^I,W^D 是彩色

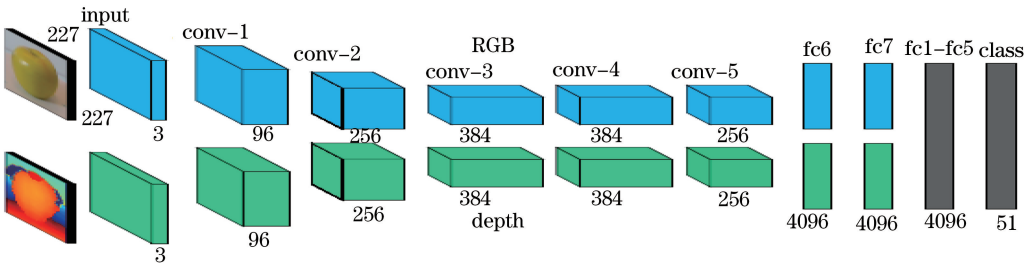

图 7 基于多模式 CNN 的稳健性 RGB-D 识别

Fig. 7 Robust RGB-D object recognition based on multimodal convolutional neural network

和深度图中全连接到输出层的权重。本文方法采用两阶段方法进行分类器模型训练：

1) 在 fc7 层，利用下面约束函数学习彩色和深度图模型，表达式为

$$\min_{w^I,\theta^I}\sum_{i=1}^{N}\ell\{\mathrm{softmax}[W^I g^I(x^i;\theta^I)],y^i\}, \tag{2}$$

$$\min_{w^D,\theta^D}\sum_{i=1}^{N}\ell\{\mathrm{softmax}[W^D g^D(d^i;\theta^D)],y^i\}, \tag{3}$$

式中 $\mathrm{softmax}(z)=\exp(z)/\|z\|_1$，$\ell(s,y)=-\sum_k y_k\log_2 s_k$。

2) 根据 fc7 层提取的彩色和深度特征，在 fc1-fus 层将这两种特征进行融合学习，令 $f\{[g^I,g^D];\theta^F\}$ 表示彩色和深度特征融合函数，通过全连接建立融合网络模型，其约束函数为

$$\min_{W^f,\theta^F,g^I,g^D}\sum_{i=1}^{N}\ell\{\mathrm{softmax}[W^f f([g^I,g^D];\theta^F)],y^i\}。 \tag{4}$$

表 3 列出该算法在 RGB-D 数据集[11]的结果(表中加粗的行)，同时列出该算法与上述特征算法分类性能比较。从表 3 每行中发现：采用自动特征学习方法(如 CNN-RNN，CNN Features，Fus-CNN)比手工特征方法(如 SIFT，Kernel Desc)具有更好的分类性能。从表 3 每列中观察到：与单独 RGB 及 Depth 图像分类性能相比，融合深度信息的 RGB-D 图像能显著提高图像的分类性能。

表 3 不同算法在 RGB-D 数据集上的分类性能比较

Table 3 Comparison of different classification methods performance in the RGB-D data set

Method	RGB	Depth	RGB-D
SIFT+SVM[23]	74.5±3.1	64.7±2.2	83.9±3.5
Kernel Desc[3]	77.7±1.9	78.8±2.7	86.2±2.1
CNN-RNN[6]	80.8±4.2	78.9±3.8	86.8±3.3
SP+HMP[1]	82.4±3.1	81.2±2.3	87.5±2.9
CNN Features[33]	83.1±2.0	N/A	89.4±1.3
Fus-CNN[54]	**84.1±2.7**	83.0± 2.7	91.0±1.9

基于 2D 特征的 RGB-D 分类方法缺乏考虑 3D 图像的位置与姿态信息，对复杂场景的多目标识别，未有较高识别精度和稳健性，只适应特定的场景，缺乏广泛实际应用。采用 3D 特征描述子，能够在多场景、多视角和目标复杂的实际数据集中获取更好的识别率，同时对遮挡噪声数据具有较强的稳健性。

4.2 基于 3D 图像特征的 RGB-D 分类

基于 3D 图像特征识别方法工作过程为：首先利用 Kinect 传感器获取深度信息，再将深度信息转换为 3D 点云模型，例如点云库(PCL)[10-11]，接着从 3D 点云模型中通过关键点匹配[34]，并提取特征描述子，最后利用这些 3D 特征进行 RGB-D 图像分类。文献[55]采用计算机图形模型(CG)渲染物体深度信息，将 RGB-D物体(例如椅子)检测分成两部分：1) 训练阶段，针对椅子在不同光照、遮挡、噪声的复杂场景下提取其三维点云的 3D 特征并建立 SVM 分类器；2) 测试阶段，完成复杂场景下椅子物体的检测识别。实现过程如图 8 所示，其在遮挡、复杂场景下检测效果如图 9 所示，RGB 中绿色盒子是遮挡的物体，从图 9 中发现，采用点云库自带深度特征描述子的目标检测性能(图 9 第四列)远远低于 CG 深度特征描述子结果(图 9 第五列)。因此，3D 特征描述子设计对 RGB-D 分类性能至关重要。

3D 特征主要有外观描述子(从彩色图像中获取)与形状特征描述子(从深度图像中获取)，研究者分别从融合外观、颜色、多视角与形状等方面，针对不同应用场合，提出很多稳健性强 3D 特征描述子。常用特征描

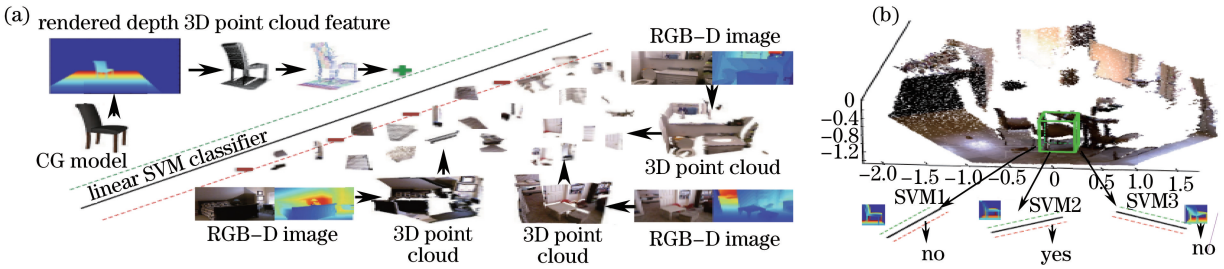

图 8 复杂场景下 3D 物体检测识别。(a) 训练；(b) 测试

Fig. 8 Detection and recognition of 3D objectsin complex scenes. (a) Training; (b) testing

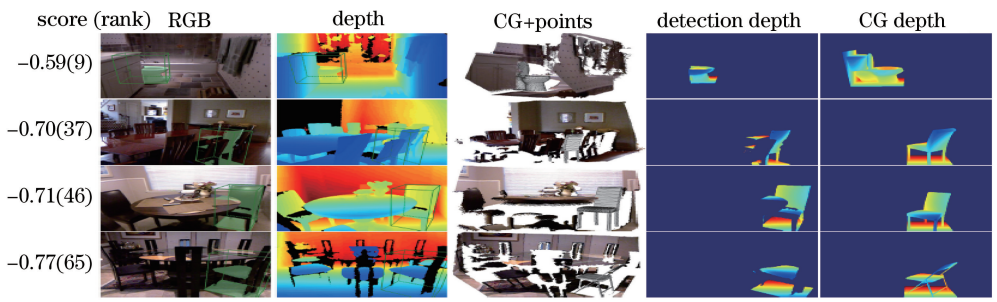

图 9 遮挡、噪声下的物体检测识别效果

Fig. 9 Object detection and recognition effect under occlusion and noise

述子如表 4 所示，主要分为两大类：全局 3D 特征描述子和局部 3D 特征描述子。全局 3D 特征描述子从分割好的图像通过聚类获取[12-14,20]，对部分遮挡效果比较敏感。相反，局部 3D 特征描述子是通过真实场景数据中逐对特征匹配提取，导致在最后识别和姿态还原中计算成本高[15-19,21]。

表 4 全局和局部 3D 描述子

Table 4 Global and local 3D descriptor

	Feature descriptor	Representation of data	Feature extraction method
Global 3D descriptor	VFH[13]	3D geometry and multi view point cloud	Histogram of combination of geometry and multi view point information
	VGH-Texton[31]	3D appearance texture point cloud	Combine appearance feature and texture shape feature
	PFHRGB[12]	3D texture color point cloud	Combination of shape and color in 3D point cloud based on color and normal
	VCSH[20]	3D color and shape histogram	Combine color and shape feature
	ESF[14]	3D appearance shape point cloud	Combine with local dotted pairs shape feature
Local 3D descriptor	CPPF[16]	Color pairs feature descriptor information	Hash table representation feature composed by geometry and HSV color information
	SHOT[17]	3D texture point cloud	Histogram of combination of point cloud appearance and shape feature
	Con VOSCH[15]	Point cloud database of vertex texture	Combination of 3D geometry information and visual RGB information feature
	BRAND[21]	RGB and depth image	Binary representation for shape and geometry information
	CSHOT[19]	3D texture point cloud	Combine shape and color histogram in 3D point cloud based on L1 norm

4.2.1 基于全局 3D 特征描述子的 RGB-D 分类

对于全局 3D 特征描述子，Rusu 等[13]提出基于点云的 3D 多视觉形状(VFH)识别对象的方法。该方法

是局部 3D 特征 FPFH[18]方法的扩展，利用融合几何与多视角特征形成直方图描述子方法，在目标表面存在大量噪声及缺少深度信息的情况下，具有很好的稳健性。实验也表明，该算法能在比较复杂和多种姿态变化的场景下实现多目标的快速识别。但这种特征描述子没有考虑物体的姿态，也没有考虑物体的纹理和颜色特征。文献[31]在文献[13]方法基础上，融合纹理形状特征，形成 VGH-Texton 特征描述子，通过外观特征与形状特征有效融合，在 3D 场景分类中具有较好的目标识别性能。文献[14]也在文献[13]方法基础上提出了全局 3D 描述子 ESF(形状特征集成)，通过在随机选择点对中融合角度，点距和面积形状因子，将局部点对特征逐渐形成全局描述子。但是这种描述特征方法忽略了物体的表面光信息，因此不能准确反映物体的姿态信息。VCSH[20]是融合颜色和形状特征构成的全局 3D 描述子，通过多视觉点云聚类物体中心点，形成几何形状特征，并产生多视角颜色特征，在多目标复杂场景中能够快速准确识别物体，在复杂的光照变化下具有稳健性。

文献[12]采用贝叶斯匹配方法进行对象识别和姿态估计，其对象全局描述子通过完整的网格模型产生，包括融合颜色、3DSIFT 纹理特征。如图 10 所示，其中红盒子对应使用点云和深度信息步骤，绿盒子对应利用全局颜色信息。蓝色盒子对应提取利用局部 SIFT 特征过程。其分类包括两个过程：训练和测试过程。在训练阶段，根据原始图像和点云数据，获取图像三维网格模型(训练过程中第 4 幅图像)；根据该图像三维网格模型、掩码信息和 SIFT 特征，获取该图像三维模型(训练过程中第 9 幅图像)；根据包含物体位置和颜色的点云数据及物体的六自由度位姿信息，产生 27 维全局颜色直方图特征。在测试阶段，根据物体颜色信息、深度信息和点云信息，首先将场景分割并聚类为各个潜在对象，然后提取这些对象的 SIFT 和全局色调直方图特征，并使用贝叶斯最近邻方法将提取的特征与训练过程获得的对象模型和全局颜色特征进行匹配，最后将匹配模型中得分最高的物体类别作为待识别物体分类结果，获得了不错的性能。然而，这种方法需要详细的网格训练模型，并且需要这些训练对象具有完整的纹理信息。采用全局 3D 特征描述子的 RGB-D 分类，因考虑了图像全局统计信息及姿态信息，而提高图像分类的稳健性。

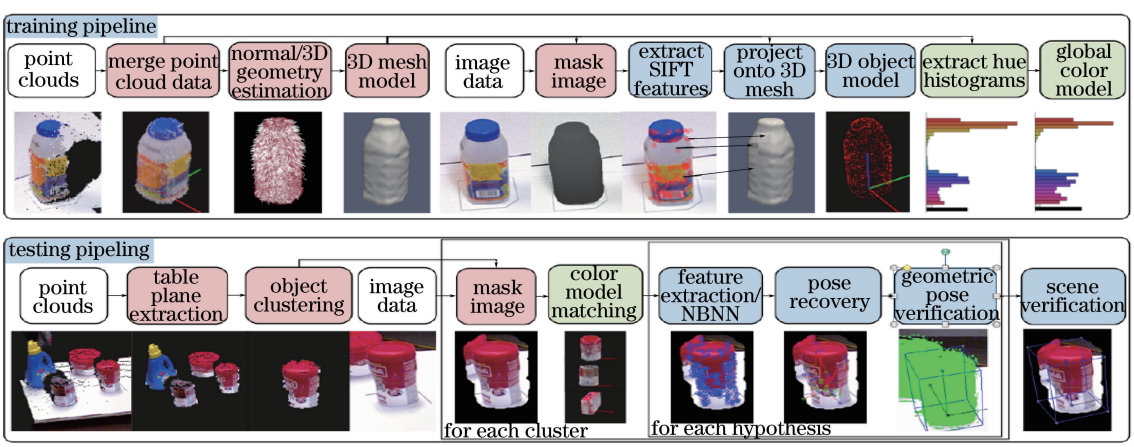

图 10 点云训练和测试过程

Fig. 10 Point clouds training and testing process

4.2.2 基于局部 3D 特征描述子的 RGB-D 图像分类

对于局部 3D 描述子，文献[17]提出 SHOT 描述子。这种特征描述根据一个输入点采用奇异值分解获得一个重复局部帧，通过这个局部帧，将球体分解成球体网格，再通过聚类方法计算该网格中心点，对每个中心点计算角度、形状并表示为直方图形成特征描述子。CSHOT 描述子[19]是在 SHOT 描述子上增加颜色信息构成，该描述子主要依靠局部参考帧，但在对象对称旋转情况下，不能稳定评估这些参考帧信息，导致识别效果不稳定。ConVOSCH 描述子[15]将 3D 几何信息同可视化 RGB 信息特征进行融合，能实时识别 3D 对象，但这种描述子不能很好地估计物体精确的姿态。Choi 等[16]提出局部对象颜色对特征描述子(CPPF)，采用融合几何与 HSV 颜色信息，利用哈希表来存储特征信息。该方法计算成本高，在不同场景中对高维参数设置很敏感。文献[21]提出基于外观与法线的稳健二进制描述子(BRAND)局部描述符，该描述符将 RGB-D 图像外观和几何形状信息进行高效融和，能保持图像旋转和缩放不变特点，在光线暗场景情况下能获得较

好的图像匹配性能,且该方法将图像点信息采用二进制字符串编码,适合于性能要求高和内存消耗低的应用。其主要的缺点是抗噪性不强,特别是对于不平的表面,几何特征很难被提取。

文献[54]采用3种关键点提取匹配方法,在RGB-D数据集[23]下,给出多种3D特征描述子在类别识别和实例识别下的识别率比较。如表5所示,从表中发现,基于3D特征的RGB-D分类方法能获得较好的类别识别结果,融合颜色及点云信息的形状特征描述子(PFHRGB[11],CHOST[19])有利于类别识别和目标识别。

表5 5种3D特征描述子的类别和实例识别

Table 5 Category and object recognition using five kinds of 3D descriptor

Descriptor	Harris 3D		Sub-sample (1 cm)		Sub-sample (2 cm)	
	Category /%	Object /%	Category /%	Object /%	Category /%	Object /%
PFHRGB[11]	93.89	77.89	94.09	79.32	94.73	79.75
ESF[14]	82.91	39.03	83.54	39.66	81.65	37.34
FPFH[18]	81.89	44.63	87.55	49.58	86.08	47.26
SHOT[17]	81.26	42.53	91.77	55.49	88.19	46.84
CSHOT[19]	88.40	69.20	92.62	75.53	90.72	73.42

5 结束语

在介绍Kinect 1.0和Kinect 2.0的结构与技术原理基础上,描述了已有的3D测试数据集特点,重点对现有的RGB-D图像目标识别和场景分类方法进行详细综述,提供较为完整的基于Kinect技术的3D物体识别的现状分析。RGB-D图像分类在以下几个方面还亟待进一步研究:

1) 深度图像修正及点云库技术研究发展。深度图像的质量直接影响着RGB-D图像分类精度。为获取遮挡或者重叠的物体整体平滑度好、边缘清晰的深度图像,需要研究修复原始深度图像中的空洞和噪声技术,同时消除深度图像中信息不一致性问题。基于深度信息的点云库不仅提供了对RGB-D信息的获取,还提供了快速的划分、特征提取、识别、寻踪等最新的算法,开发完善功能强大的点云库对RGB-D分类至关重要。

2) 物体局部与全局特征自动学习及融合策略研究。利用深度学习模型自动学习RGB-D图像的局部和全局特征是未来的发展趋势。物体局部特征考虑物体本身结构信息,使得物体分类的准确性更高,但同时也带来物体分类的稳健性不强的问题;全局特征考虑更多的是图像全局统计信息,尤其是图像的语义信息,使其稳健性能够得到一定地提高。目前RGB-D分类分别从彩色图像和深度图像提取特征,没有充分考虑它们之间的相互作用,是单一模式,今后将彩色与深度图像同时进行层次特征融合,挖掘它们之间的相互联系,发展多种模式多特征学习。

3) 与基于2D图像特征分类方法比较,基于3D图像特征识别算法有着更广阔的发展前景。随着图形处理器(GPU)硬件和并行计算的发展,深度信息向3D点云模型转换及匹配的计算速度越来越快,且3D图像具有更好的精确度和稳健性。3D图像能够提供许多强有力的目标线索,如:目标大小、形状、方向及边界等信息,将显著提升基于3D图像特征的RGB-D分类性能,促进其在各个领域中广泛应用。

参 考 文 献

1 Bo L, Ren X, Fox D. Unsupervised feature learning for RGB-D based object recognition[J]. Springer Tracts in Advanced Robotics, 2013, 88: 387-402.

2 Lai K, Bo L, Ren X, *et al*.. Sparse distance learning for object recognition combining RGB and depth information[C]. Robotics and Automation, International Conference on IEEE, 2011: 4007-4013.

3 Bo L, Ren X, Fox D. Depth kernel descriptors for object recognition[C]. Intelligent Robots and Systems (IROS), International Conference on IEEE, 2011: 821-826.

4 Bium M, Springenberg J T, Wulfing J, *et al*.. A learned feature descriptor for object recognition in RGB-D data[C]. Proceedings of IEEE International Conference on Robotics and Automation, 2012: 1298-1303.

5 Kramer J, Burrus N, Echtler F, *et al*.. Hardware[J]. Hacking the Kinect, 2012, 14(2): 156-156.

6 Socher R, Huval B, Bath B, *et al*.. Convolutional-recursive deep learning for 3d object classification[C]. Advances in

Neural Information Processing Systems, 2012: 665-673.

7 Couprie C, Farabet C, Najman L, *et al*.. Indoor semantic segmentation using depth information[C]. International Conference on Learning Representations, Scottsdale, Arizona, 2013.

8 Farabet C, Couprie C, Najamn L, *et al*.. Learning hierarchical features for scene labeling[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8): 1915-1929.

9 Gupta S, Girshick R, Pablo A, *et al*.. Learning rich features from RGB-D images for object detection and segmentation [C]. European Conference on Computer Vision, Zurich, Switzerland, 2014: 345-360.

10 Rusu R B, Cousins S. 3D is here: Point cloud library (PCL)[C]. Robotics and Automation (ICRA), International Conference on IEEE, 2011: 1-4.

11 Rusu R B, Blodow N, Marton Z C, *et al*.. Aligning point cloud views using persistent feature histograms[C]. Intelligent Robots and Systems, IROS 2008, IEEE/RSJ International Conference on IEEE, 2008: 3384-391.

12 Tang J, Miller S, Singh A, *et al*.. A textured object recognition pipeline for color and depth image data[C]. Robotics and Automation (ICRA), 2012 IEEE International Conference on IEEE, 2012: 3467-3474.

13 Rusu R B, Bradski G, Thibaux R, *et al*.. Fast 3D recognition and pose using the viewpoint feature histogram[C]. Intelligent Robots and Systems (IROS), International Conference on IEEE, 2010: 2155-2162.

14 Wohlkinger W, Vincze M. Ensemble of shape functions for 3d object classification[C]. 2011 IEEE International Conference on Robotics and Biomimetics, 2011: 2987-2992.

15 Kanezak A, Marton Z, Pangercic D, *et al*.. Voxelized shape and color histograms for RGB-D[C]. IROS Workshop on Active Semantic Perception, 2011.

16 Choi C, Christensen H I. 3D pose estimation of daily objects using an RGB-D camera[C]. Intelligent Robots and Systems (IROS), International Conference on IEEE, 2012: 3342-3349.

17 Tombari F, Salti S, Stefano L D. Acombined texture-shape descriptor for enhanced 3D feature matching[C]. Image Processing (ICIP), 2011 18th IEEE International Conference on 2011, 2011: 809-812.

18 Rusu R B, Blodow N, Beetz M. Fast point feature histograms (FPFH) for 3D registration[C]. Proceedings of the IEEE international conference on Robotics and Automation IEEE Press, 2009: 3212-3217.

19 Wohlkinger W, Vincze M. Ensemble of shape functions for 3d object classification[C]. IEEE International Conference on Robotics and Biomimetics (ROBIO), 2011: 2987-2992.

20 Wang W, Chen L, Liu Z, *et al*.. Textured/textureless object recognition and pose estimation using RGB-D image[J]. Journal of Real-Time Image Processing, 2013: 1-16.

21 Nascimento E R, Oliveira G L, Campos M F M, *et al*.. BRAND: A robust appearance and depth descriptor for RGB-D images[C]. Intelligent Robots and Systems (IROS), International Conference on IEEE, 2012: 1720-1726.

22 Gupta S, Arbelaez P, Malik J. Perceptualorganization and recognition of indoor scenes from RGB-D images[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013: 564-571.

23 Lai K, B O L, REN X, *et al*.. A large-scale hierarchical multi-view RGB-D object dataset[C]. Robotics and Automation (ICRA), International Conference on IEEE, 2011: 1817-1824.

24 Janoch A, Karayev S, Jia Y, *et al*.. A category-level 3d object dataset: Putting the Kinect to work[M]. London: Springer, 2013: 141-165.

25 Silberman N, Hoiem D, Kohli P, *et al*.. Indoor segmentation and support inference from RGBD images[M]. Heidelberg: Springer, 2012: 746-760.

26 Hema S K, Abhishek A, Joachims T, *et al*.. Semantic labeling of 3D point clouds for indoor scenes[J]. Nips, 2011: 244-252.

27 Xiao J, Owens A, TorralbaA. SUN3D: Adatabase of big spaces reconstructed using SfM and object labels[C]. IEEE International Conference on Computer VisionInstitute of Electrical and Electronics Engineers, 2014: 1625-1632.

28 Song S, Lichtenberg S P, Xiao J. Sun RGB-D: A RGB-D scene understanding benchmark suite[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 567-576.

29 Fanelli G, Dantone M, Gall J, *et al*.. Random forests for real time 3D face analysis[J]. International Journal of Computer Vision, 2013, 101(3): 437-458.

30 Hinterstoisser S, Lepetit V, Ilic S, *et al*.. Model based training, detection and pose estimation of texture-less 3D objects

in heavily cluttered scenes[J]. Lecture Notes in Computer Science, 2012.

31 Ali H, Shafait F, Giannakidou E, *et al*.. Contextual object category recognition for RGB-D scene labeling[J]. Robotics & Autonomous Systems, 2014, 62(2): 241-256.

32 Yang C, Jang Y, Beh J, *et al*.. Gesture recognition using depth-based hand tracking for contactless controller application [C]. Digest of Technical Papers-IEEE International Conference on Consumer Electronics, 2012: 297-298.

33 Schwarz M, Schulz H, Behnke S. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features[C]. IEEE International Conference on Robotics & Automation, 2015.

34 Zhou Wei, Liu Gang, Ma Xiaodan, *et al*.. Study on multi-image registration of apple tree at different growth stages[J]. Acta Optica Sinica, 2014, 34(2): 0215001.
周 薇, 刘 刚, 马晓丹, 等. 不同生长时期果树多源图像的配准方法研究[J]. 光学学报, 2014, 34(2): 0215001.

35 Koppula H S, Gupta R, Saxena A. Learning human activities and object affordances from RGB-D videos[J]. International Journal of Robotics Research, 2012, 32(8): 951-970.

36 Ni B, Wang G, Mouli P. Rgbd-hudaact: A color-depth video database for human daily activity recognition [C]. In Consumer Depth Cameras for Computer Vision, 2013: 193-208.

37 Li Xiuzhi, Yang Ailin, Qin Baoling, *et al*.. Monocular camera three dimensional reconstruction based on optical flow feedback[J]. Acta Optica Sinica, 2015, 35(5): 0515001.
李秀智, 杨爱林, 秦宝岭, 等. 基于光流反馈的单目视觉三维重建[J]. 光学学报, 2015, 35(5): 0515001.

38 Jia Songmin, Wang Ke, Li Xiuzhi, *et al*.. Monocular camera three dimensional reconstruction based on variation model [J]. Acta Optica Sinica, 2014, 34(4): 0415002.
贾松敏, 王 可, 李秀智, 等. 基于变分模型的单目视觉三维重建方法[J]. 光学学报, 2014, 34(4): 0415002.

39 Tu S Q, Xue Y J, Liang Y, *et al*.. Learning structured group sparse representation for RGB-D image classification[J]. Journal of Information and Computational Science, 2015, 12(11): 4357-4367.

40 Huang Xiaolin, Xue Yueju, Tu Shuqin, *et al*.. RGB-D images classification based on compressed sensing theory[J]. Computer Applications and Software, 2014, 31(3): 195-197.
黄晓琳, 薛月菊, 涂淑琴, 等. 基于压缩感知理论的 RGB-D 图像分类方法[J]. 计算机应用与软件, 2014, 31(3): 195-197.

41 Handa A, Whelan T, Mcdonald J, *et al*.. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM[C]. Robotics and Automation (ICRA), International Conference on IEEE, 2014: 1524-1531.

42 Burgard W, Cremers D, Sturm J, *et al*.. A benchmark for the evaluation of RGB-D SLAM systems[C]. International Conference on Intelligent Robot Systems, 2012: 573-580.

43 Shotton J, Girshick R, Fitzgibbon A, *et al*.. Efficient human pose estimation from single depth images[M]. London: Springer, 2013: 175-192.

44 Hinterstoisser S, Lepetit V, Ilic S, *et al*.. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[J]. Lecture Notes in Computer Science, 2012.

45 Wang W, Chen L, Liu Z, *et al*.. Textured/textureless object recognition and pose estimation using RGB-D image[J]. Journal of Real-Time Image Processing, 2013: 1-16.

46 Ohn-Bar E, Trivedi M M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(6): 2368-2377.

47 Yin Panlong, Xu Guangzhu, Lei Bangjun, *et al*.. Review on the technique to obtain depth information using Kinect and its application to three dimensional object recognition[J]. Journal of Integration Technology, 2013: 2(6): 94-99.
尹潘龙, 徐光柱, 雷帮军, 等. Kinect 下深度信息获取技术及其在三维目标识别中的应用综述[J]. 集成技术, 2013: 2(6): 94-99.

48 Quigley M, Conley K, Gerkey B, *et al*.. ROS: An open-source robot operating system[C]. ICRA Workshop on Open Source Software, 2009, 3(3.2): 5.

49 Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor[C]. Computer Vision Workshops (ICCV Workshops), International Conference on IEEE, 2011: 601-608.

50 Ren X, Bo L, Fox D. RGB-(D) scene labeling: Features and algorithm[C]. Computer Vision and Pattern Recognition, IEEE, 2012: 2759-2766.

51 Cheng Y, Zhao X, Huang K, *et al*.. Semi-supervised learning for RGB-D object recognition[C]. Pattern Recognition, International Conference on IEEE, 2014.

52 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(8): 1798-1828.

53 Wang A, Lu J, Wang G, *et al*.. Multi-modal unsupervised feature learning for RGB-D scene labeling[M]. London: Springer, 2014: 453-467.

54 Eitel A, Springenberg J T, Spinello L, *et al*.. Multimodal deep learning for robust RGB-D object recognition[C]. CVPR, 2015.

55 Song S, Xiao J. Sliding shapes for 3D object detection in depth images[M]. London: Springer, 2014: 634-651.