文章编号:1007-1423(2021)14-0092-04 DOI:10.3969/j.issn.1007-1423.2021.14.018

基于 HoloLens2 的目标检测技术研究

曾笑

(四川大学视觉合成图形图像技术国防重点学科实验室,成都 610065)

摘要:

利用深度学习的目标检测技术与混合现实结合的方法,旨在帮助穿戴 Microsoft HoloLens2 设备的用户对真实世界的物体对象进行实时高效的检测,得到物体对象的位置,还能够识别目标,得到目标的分类信息及其所属类别的置信度。同时,利用 HoloLens2 的增强现实信息能够更加直观地看到检测信息,摆脱传统的人机交互方式,达到虚实交互的目的。实验结果表明,该系统能够较快地计算目标检测信息,具有较高的识别准确率。

关键词:

目标检测; Microsoft HoloLens2; 混合现实; 深度学习

0 引言

计算机视觉的任务是为了让计算机能够像人眼一样理解图像里的内容,而目标检测一直是计算机视觉的热门内容。目标检测技术是通过算法判断图片中需要检测的物体,同时在图片中标记出该物体的位置,用计算出的边框将其圈起来,并返回分类类别。传统的目标检测技术是通过人工特征描述算子对图像进行描述后,再依据描述特征来对图像进行查找[1]。而近年来,随着深度学习在计算机视觉上的发展,利用卷积神经网络的相关知识在解决如目标检测任务上也开始普及了起来,在识别准确率和鲁棒性上都优于传统检测算法。然而与目标检测相关的主流交互设备仍然是传统的键盘、鼠标和触摸屏,这些设备在进行目标检测与人机交互方面具有很大的局限性。

混合现实技术是一种将虚拟对象与真实世界巧妙融合的技术。虚拟对象与真实世界通过计算机的实时计算,可以实现无缝叠加,从而达到虚实融合的目的,给用户带来极强的真实感受。而混合现实技术的发展,可以将应用场景应用在教育、工业、医疗等行业。目前大多数的领域还停留在探索和发展的阶段。

Microsoft HoloLens2 是微软在 2019 年发布的第二代混合现实眼镜,其处理器使用了一块 Intel 32bit CPU 和一个定制的高性能混合现实运算单元 HPU(Holographic Processing Unit)。相比于其他 AR 设备,HoloLens2 采用新的交互模式和三维注册算法,不需要额外辅助定位的器件,就能够计算虚拟对象与现实场景之间的空间位置关系,将物理世界与数字世界相融合,完全解放双手,实现虚实叠加、人机交互等技术。通过全息影像的方式,将数字内容展现出来,同时还可以通过注视、手势、语音来与全息影像进行交互。

通过将混合现实以及目标检测的相关技术进行整合,能够提供一种新的人机交互方式,为学习培训、可视化展示等行业所存在的问题提供了理论与技术上的支持,从而使得设计并开发出可交付的解决方案成为可能。

在本文工作中,我们开发了一个目标检测系统,通过通信的方式,能够在用户端使用 Microsoft HoloLens2 对目标进行检测和识别,在服务器端通过 YOLO 作为深度学习算法处理来自客户端的数据。客户端与服务端之间通过 TCP/IP 进行通信,服务器端可以处理需要进行检测的对象,并将处理过的数据传回客户端。本文主要分为以下几个部分,第一节介绍了基于深度学习的目标检测算法,第二节介绍了我们设计的目标检测系统,具体描述了目标检测算法是如何在 HoloLens2

上进行实现的。第三节对于我们的实验结果进行分析与讨论。最后第四节提出我们所做工作的总结以及对于未来的展望。

1 基于YOLO卷积神经网络的目标检测

YOLO(You Only Look Once)是 Redmon 提出的目标检测算法[2],它基于一个单独的 end-to-end 网络,将物体分类和位置识别全部统一到神经网络的一次检测当中。输入图像经过网络,便能得到图像中所有物体的位置和其所属类别及相应的置信概率,使得检测速度大幅提升。虽然 YOLOv1 在当时成功实现了实时的目标检测,但是图像中包含多个重叠物体时还是会出现偏差,同时在复杂的环境中还易出现边界框不准的情况。

针对上述问题,在 2017 年,Redmon 在 YOLOv1 的基础上提出 YOLOv2 和 YOLO9000[3],能够检测超过 9000 类物体。与 YOLOv1 相比,它加入了 Batch Normalization[4],能够提升模型收敛速度,降低模型的过拟合;同时能够接受不同分辨率的输入图片,引入不同尺度特征融合,从而提高了训练速度和分类效果。

紧接着在 2018 年,Redmon 提出了 YOLOv3[5]。相比于 YOLOv2,YOLOv3 具有以下几个变化:调整了新的网络结构,使用了残差模块[6]的 DarkNet-53 模型;采用 FPN 结构实现多尺度特征提取,实现目标检测;对象分类用 Logistic 取代了 Softmax。通过形成更深的网络层次以及多尺度检测,提升了 mAP 及物体的检测效果。在 YOLOv3 中定义的损失函数由三个部分组成,具体公式如下所示:

$$lbox = \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{obj}(2-w_i*h_i)[(x_i-\hat{x_i})^2+(y_i-\hat{y_i})^2+(w_i-\hat{w_i})^2+(h_i-\hat{h_i})^2] \quad (1)$$

$$lcls = \lambda_{class}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{obj}\sum_{c\in classes}p_i(c)\log(\hat{p_i}(c)) \quad (2)$$

$$lobj = \lambda_{noobj}\sum_{i=0}^{S^2}1_{i,j}^{obj}(c_i-\hat{c_i})^2+\lambda_{obj}\sum_{i=0}^{S^2}1_{i,j}^{obj}(c_i-\hat{c_i})^2 \quad (3)$$

$$loss = lbox + lobj + lcls \quad (4)$$

其中,*lbox* 代表 bounding-box 带来的 *loss*;*lobj* 是置信度带来的 *loss*,置信度代表该 bounding-box 是否含有物体的概率;*lcls* 则是类别带来的 *loss*。通过实验比较可以得知,在精确度相当的情况下,YOLOv3 的速度约是其他模型的 3、4 倍,在准确性和实时性上均满足系统要求,因此本文采取 YOLOv3 神经网络来实现对物体的检测识别定位。

2 系统架构设计

2.1 总体设计

这一部分我们将描述所设计的系统整体架构,目的是为了使用目标检测算法来对图像中的物体进行识别检测。我们主要采用了客户端与服务器端进行通信的方式来实现交互。我们使用 Microsoft HoloLens2 作为图像采集的输入,使用具有 NVIDIA GeForce GTX 1070 Ti 的图形处理单元的电脑作为服务器端,该服务器端用于执行目标检测算法。开发人员能将客户端摄像头所采集到的图像帧传输到服务器端,作为其算法输入,算法执行输出该物体的分类类别,并通过通信返回到 HoloLens2 作为输出显示。同时,我们导入微软的 Mixed Reality Toolkit 软件开发包来控制凝视、手势识别、虚实交互。整个混合现实和深度学习的总体设计如图 1 所示。

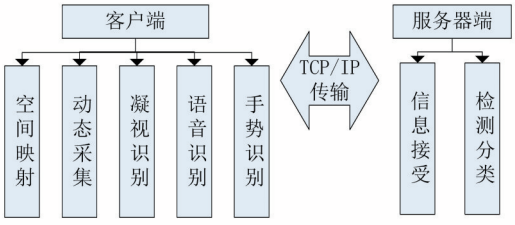

图1 系统总体设计图

2.2 客户端

在客户端创建阶段,利用 Unity 创建了项目,通过导入 Mixed Reality Toolkit 组件在 HoloLens2 上进行构建和部署独立的应用程序。在检测过程中,应用空间映射步骤,先用 HoloLens2 的相机对物理环境进行解析,用户可以通过手势指令或者语音识别指令动态采集真实环境中的照片,并通过客户端将其上传到服务器端。在服务器端运行后,快速返回检测对象的信息、分类类别以及置信度,并通过网络传输返回到客户端。服务器端将识别到的 2D 坐标返回给 HoloLens2,然后根据接收到的 2D 数据在设备上计算物体的 3D 真实坐标。最后通过生成全息图像将以上分类信息投射到用户眼睛,实时显示虚拟场景,从而实现高效的目

标检测和人机交互。

图2 第二代HoloLens头盔

2.3 服务器端

在服务器端，我们使用YOLO算法来对环境中的对象进行检测和识别。当作为客户端的HoloLens2摄像头进行拍照后，通过连接服务器IP地址的形式，将得到的RGB帧传到服务器端，作为输入。在得到接受帧后，服务器端启用YOLO神经网络对其进行处理，从而得到所检测到每个对象的分类、边界框范围以及框的颜色。完成对象检测后，服务器将这些结果发送给客户端，从而在头盔上显示检测结果。

表1 服务器端开发工具

Operating system	Windows 10
CPU	Intel Core i7-8700K CPU @ 3.70GHz
GPU	NVIDIA GeForce GTX 1070 Ti
CUDA	V10.0.130
CUDNN	V7.6.4
Visual studio	Visual studio 2019
Unity	V2019.3.15f1
Algorithm	YOLOv3

3 实验结果和讨论

本文通过HoloLens2实现目标检测和识别，取得了良好的检测效果。为了更好地判断检测效果，在实验结果中，通过召回率(Recall)以及精确率(Precision)来判别检测结果的好坏，其中，召回率表示返回的正样本中预测正确的样本数，占真正总的正样本数的比例；精确率表示计算返回的正样本中，预测正确的比例。同时再对比算法在测试集上的平均准确率(AP)以及所有类别的平均准确率(mAP)来判别算法的好坏。其中AP需要通过召回率和准确率两个值来衡量检测模型的精确度，它将用作评估模型检测性能的直观标准，同时也可以用来评估单个类别的准确度。mAP是所有AP类别的平均值，值越高，代表在其所有类别中的模型检测效果越好。整个计算过程如下：

(1)先按照置信度将n个测试结果按从大到小的方式进行排序；

(2)定义TP数组表示正类预测为正的例子，FN数组表示被预测为负值的正样例，FP数组表示预测为正的负样例。

(3)召回率、精确率分别用下列公式得出：

$$R[i]=\frac{\sum_{j=1}^{i}TP[j]}{\sum_{j=1}^{i}TP[j]+\sum_{j=1}^{i}FN[j]},j=1,2,\cdots,n \tag{5}$$

$$P[i]=\frac{\sum_{j=1}^{i}TP[j]}{\sum_{j=1}^{i}TP[j]+\sum_{j=1}^{i}FP[j]},j=1,2,\cdots,n \tag{6}$$

(4)计算某类别的AP：

$$AP_i=\frac{1}{n}\sum_{r\in\{0,0.1,\cdots,1\}}^{i}\max_{\tilde{r}:\tilde{r}\geq r}P(\tilde{r}),i=1,2,\cdots,n \tag{7}$$

(5)计算所有类别的AP均值，即mAP：

$$mAP=\frac{1}{n}\sum_{i=1}^{n}AP_i \tag{8}$$

最终训练后得到的各类别的Recall和Precision的结果如表2所示。

表2 各对象的召回率和精确率

分类	召回率(Recall)	精确率(Precision)
Bottle	73.6%	97.7%
Chair	72.1%	97.6%
Mouse	61.4%	98.8%
Monitor	85.4%	94.5%
Cup	71.9%	98.8%
Laptop	76.4%	93.9%

4 结语

本文通过将最新的增强现实设备HoloLens2和先进的目标检测算法YOLOv3结合起来，通过TCP/IP连接在客户端和服务器端建立了联系，设计并实现了支持对象检测、可视化人机交互的增强现实系统。通过实验结果表明，本文提出的方法在目标检测上具有高召回率和高精确率，在测试集上的准确率和所有类别的平均准确率都具有较好的表现。同时，利用第二代HoloLens增强现实头盔，摆脱了传统的人机交互手段，可以解放双手，支持多种交互手段，为提供可视化展示提供了理论和技术支持。但本文在设计实验时缺少对照实验，且对算法本身的改动较少，在未来的工作中会继续完善实验。

参考文献：
[1]DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005: 886-893.
[2]REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.
[3]REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.
[4]]IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[J]. arXiv Preprint arXiv: 1502.03167, 2015.
[5]REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv Preprint arXiv: 1804.02767, 2018.
[6]HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

作者简介：
曾笑(1996-)，男，硕士研究生，研究方向为计算机图形图像与虚拟现实
收稿日期：2021-01-19　　修稿日期：2021-03-27

Research on Object Detection Technology Based on HoloLens2

ZENG Xiao

(Sichuan University National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065)

Abstract:
The method of combining deep learning object detection technology and mixed reality is designed to help users wearing Microsoft HoloLens2 devices to detect real-world objects in real-time and efficiently, to obtain the position of the object, and to recognize the target and get the target classification The confidence level of the information and its category. At the same time, using the augmented reality information of HoloLens2 can see the detection information more intuitively, get rid of the traditional human-computer interaction method, and achieve the purpose of virtual and real interaction. Experimental results show that the system can calculate target detection information faster and has a higher recognition accuracy.

Keywords:
Object Detection; Microsoft HoloLens2; Mixed Reality; Deep Learning