基于深度学习的多视窗SSD 目标检测方法

唐 聪[1,2,3],凌永顺[1,2,3],郑科栋[4],杨 星[1,3],郑 超[1,2,3],杨 华[1,2,3],金 伟[1,2,3]

(1. 国防科技大学,安徽 合肥 230037; 2. 红外与低温等离子体安徽省重点实验室,安徽 合肥 230037;
3. 脉冲功率激光技术国家重点实验室,安徽 合肥 230037;
4. 中国人民解放军 31101 部队,江苏 南京 210018)

摘 要:提出了一种基于深度学习的多视窗 SSD(Single Shot multibox Detector)目标检测方法。首先阐述了经典 SSD 方法的模型与工作原理,并根据卷积感受野的概念和模型特征层与原始图像的映射关系,分析了各层级卷积感受野大小和特征层上默认框在原始图像上的映射区域尺寸,揭示了经典 SSD 方法在小目标检测上不足的原因。基于此,提出了一种多视窗 SSD 模型,阐述了其模型结构与工作原理,并通过 106 张小目标图像数据集测试,评估和对比了多视窗 SSD 方法与经典 SSD 方法在小目标检测上的物体检索能力与物体检测精度。结果表明:在置信度阈值为 0.4 的条件下,多视窗 SSD 方法的 AF(Average F-measure)为 0.729,mAP(mean Average Precision)为 0.644,相比于经典 SSD 方法分别提高了 0.169 和 0.131,验证了所提出算法的有效性。

关键词:深度学习; 多视窗 SSD; 目标检测; 小目标

中图分类号:TP391.4 **文献标志码**:A **DOI**:10.3788/IRLA201847.0126003

Object detection method of multi-view SSD based on deep learning

Tang Cong[1,2,3], Ling Yongshun[1,2,3], Zheng Kedong[4], Yang Xing[1,3], Zheng Chao[1,2,3], Yang Hua[1,2,3], Jin Wei[1,2,3]

(1. National University of Defense Technology, Hefei 230037, China;
2. Key Laboratory of Infrared and Low Temperature Plasma of Anhui Province, Hefei 230037, China;
3. State Key Laboratory of Pulsed Power Laser Technology, Hefei 230037, China;
4. 31101 Troops of PLA, Nanjing 210018, China)

Abstract: The object detection method of multi-view Single Shot multibox Detector(SSD) based on deep learning was proposed. Firstly, the model and the working principle of classical SSD were expounded. According to the concept of convolution receptive field and the mapping relationship between the feature map and the original image, the sizes of covolution receptive field in different levels and the scales of the default boxes mapped to the original image were analyzed to find the reason why the classical SSD was not good at small object detection. Based on this, the multi-view SSD model was put forward, and the model architecture and its working principle were deeply expounded. Then, through the test in a dataset of 106 images for small object detection, the detection performance of multi-view SSD and

收稿日期:2017-06-11; **修订日期**:2017-08-12

基金项目:国家自然科学基金(61503394,61405248);安徽省自然科学基金(1508085QF121)

作者简介:唐聪(1989-),男,博士生,主要从事计算机视觉、深度学习、模式识别等方面的研究。Email:tangcong_eei@163.com

导师简介:凌永顺(1937-),男,中国工程院院士,教授,博士生导师,主要从事光电工程等方面的研究。Email:lys@126.com

classical SSD were evaluated and compared in object retrieval ability and object detection precision. Experimental results show that with the confidence threshold of 0.4, the multi-view SSD is 0.729 in Average F-measure(AF) and 0.644 in mean Average Precision(mAP), and has respectively raised 0.169 and 0.131 compared to the classical SSD in the two evaluation indexes, thus verifying the effectiveness of the proposed method.

Key words: deep learning; multi-view SSD; object detection; small object

0 引 言

目标检测已经成为计算机视觉领域重要的研究方向和研究热点[1],可应用于无人驾驶、机器人、视频监控、行人检测、海面舰船检测等领域[2-4]。在深度学习出现以前，目标检测方法主要是根据一定的先验知识,通过建立某种数学模型来完成,应用较广泛的方法有:Hough 变换[5]、帧差法[6]、背景减除法[7]、光流法[8]、滑动窗口模型[9]、可变形部件模型[10]等。具体地说，前四种方法主要采用特征+数学模型的模式,利用数据某种特性的特征来建立数学模型,求解模型得到目标检测的结果；后两种方法则主要采用特征提取+分类的模式,利用手工设计特征(如 SHIFT[11]、HOG[12]、Haar[13])并结合分类器(如 SVM[14]、Adaboost[15]),根据特征进行分类得到目标检测结果。近年来，深度学习技术的出现革新了目标检测的模式,提升了目标检测的精度和鲁棒性。基于深度学习的目标检测模型，由于深度神经网络能够自主学习不同层级的特征,相比于传统手工设计特征,学习的特征更丰富,特征表达能力更强[16]。

目前，基于深度学习的目标检测方法主要分为两类:基于区域候选的模型和基于回归的模型。基于区域候选的深度学习目标检测模型建立在区域候选的思想上,首先对检测区域提取候选区域,为后续特征提取和分类做准备,典型代表为:R-CNN[17]、SPP-net[18]、Fast R-CNN[19]、Faster R-CNN[20]、R-FCN[21]。基于回归的深度学习目标检测模型则采用回归的思想,需要预先按照一定方式划定默认框,从而建立起预测框、默认框、ground truth 物体框的关系以进行训练,典型代表为:YOLO[22]、SSD[23]。在上述几种算法中,SSD 的检测性能相对更好，同时具有可实时、准确度高两个优点,其在单块 Nvidia Titan X 显卡上检测速度可达到 58 fps (图片尺寸为 300×300),在 VOC2007 测试集上 mAP 达到 0.721。但是,SSD 对小目标的检测性能却令人不甚满意[23]。

文中提出了一种基于深度学习的多视窗 SSD 目标检测方法,主要用于改进经典 SSD 在小目标检测上的不足。该方法以多视窗为出发点,将融合所设定的五个视窗的目标检测结果，以提升目标检测能力。实验结果表明,该设计大大改善了对小目标的检测能力，这对于深度学习技术进一步应用于目标检测具有重要意义和参考价值。

1 SSD 目标检测

SSD 是一种单次检测深度神经网络，同时结合了 YOLO 的回归思想和 Faster R-CNN 的 anchors 机制[20]。采用回归的思想可以简化神经网络的计算复杂度,提高算法的实时性;采用 anchors 机制可以提取不同宽高比尺寸的特征,同时,这种局部特征提取的方法在识别方面，相比于 YOLO 针对某一位置进行全局特征提取的方法更合理、有效。另外,SSD 针对不同尺度的特征表达不同这一特点，采取了多尺度[24]目标特征提取的方法,该设计有助于提升检测不同尺度目标的鲁棒性。

1.1 SSD 模型

SSD 的架构主要分为两部分:一部分是位于前端的深度卷积神经网络,采用的是去除分类层的图像分类网络,如 VGG[25],用于目标初步特征提取;另一部分是位于后端的多尺度特征检测网络,是一组级联的卷积神经网络,将前端网络产生的特征层进行不同尺度条件下的特征提取。SSD 框架如图 1 所示。

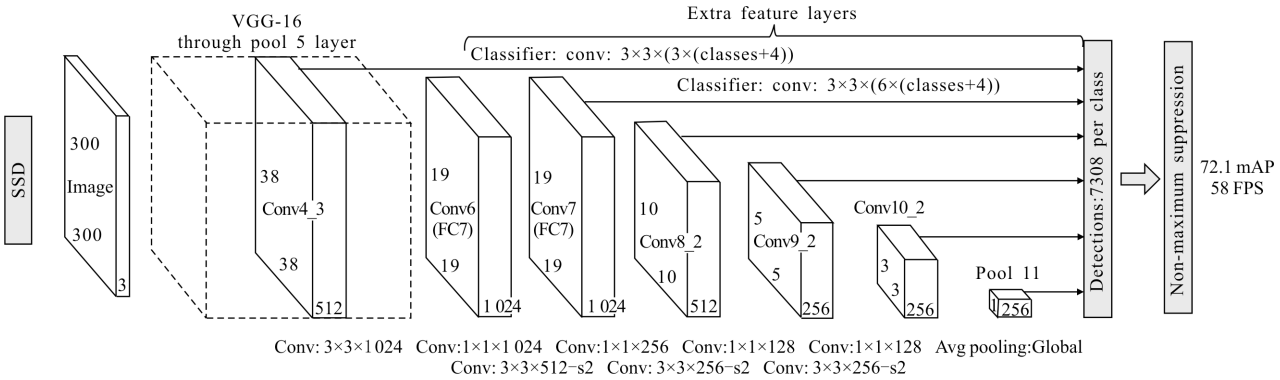

图 1 SSD 框架

Fig.1 Framework of SSD

1.2 特征层默认框映射

SSD 采用多尺度的方法可得到多个不同尺寸的特征图,假设模型检测时采用 m 层特征图,则第 k 个特征图的默认框比例计算公式如下:

$$s_k=s_{min}+\frac{s_{max}-s_{min}}{m-1}(k-1),k\in\{1,2,\cdots,m\} \tag{1}$$

式中:s_{min} 一般取 0.2;s_{max} 一般取 0.95,分别代表特征层默认框占输入图像的比例。

同时,SSD 采用了 anchors 机制,对于同一特征层上的默认框采取不同的宽高比,以增强默认框对物体形状的鲁棒性。这里,默认框的宽高比采用 $r=\{1,2,1/2,3,1/3\}$,同时,针对宽高比等于 1 这一类,添加 $s_k'=\sqrt{s_ks_{k+1}}$,则:

$$w_k^n=s_k\sqrt{r_n},h_k^n=s_k/\sqrt{r_n},n\in\{1,2,3,4,5\} \tag{2}$$

$$w_k^6=h_k^6=\sqrt{s_ks_{k+1}} \tag{3}$$

设定默认框的中心为 $\left(\frac{a+0.5}{|f_k|},\frac{b+0.5}{|f_k|}\right)$,其中,$|f_k|$ 是第 k 个特征图的尺寸大小,$a,b\in\{0,1,2,\cdots,|f_k|-1\}$,并截取默认框的坐标使其在[0,1]内。

特征图上默认框坐标与原始图像坐标的映射关系如下:

$$x_{min}=\frac{c_x+\frac{w_b}{2}}{w_{feature}}w_{img}=\left(\frac{a+0.5}{|f_k|}-\frac{w_k}{2}\right)w_{img} \tag{4}$$

$$y_{min}=\frac{c_y+\frac{h_b}{2}}{h_{feature}}h_{img}=\left(\frac{b+0.5}{|f_k|}-\frac{h_k}{2}\right)h_{img} \tag{5}$$

$$x_{max}=\frac{c_x+\frac{w_b}{2}}{w_{feature}}w_{img}=\left(\frac{a+0.5}{|f_k|}+\frac{w_k}{2}\right)w_{img} \tag{6}$$

$$y_{max}=\frac{c_y+\frac{h_b}{2}}{h_{feature}}h_{img}=\left(\frac{b+0.5}{|f_k|}+\frac{h_k}{2}\right)h_{img} \tag{7}$$

式中:(c_x,c_y)为特征层上默认框中心的坐标;w_b,h_b 为默认框的宽和高;$w_{feature},h_{feature}$ 为特征层的宽和高;w_{img},h_{img} 为原始图像的宽和高。求得的$(x_{min},y_{min},x_{max},y_{max})$为第 k 层特征图上中心为$\left(\frac{a+0.5}{|f_k|},\frac{b+0.5}{|f_k|}\right)$、大小为 w_k,h_k 的默认框映射到原始图像的物体框坐标。

1.3 损失函数

SSD 的训练同时对位置和目标种类进行回归,其目标损失函数是置信损失和位置损失之和,其表达式如下:

$$L(z,c,l,g)=\frac{1}{N}(L_{conf}(z,c)+\alpha L_{loc}(z,l,g)) \tag{8}$$

式中:N 是与 ground truth 物体框匹配的默认框个数;$L_{conf}(z,c)$为置信损失;$L_{loc}(z,l,g)$是位置损失,这里采用的是 Smooth L1 Loss[17];z 为默认框与不同类别的 ground truth 物体框的匹配结果;c 为预测物体框的置信度;l 为预测物体框的位置信息;g 为 ground truth 物体框的位置信息;α 为权衡置信损失和位置损失的参数,一般设置为 1。

该目标损失函数同时包含置信损失和位置损失,在训练过程中,通过减小损失函数值可以确保在提升预测框类别置信度的同时也提高预测框的位置可信度,而用于数据集训练,通过多次结果优化,不断提高模型的目标检测性能,从而训练出性能较好的预测模型。

2 多视窗 SSD 模型

2.1 SSD 在目标检测中的不足

由于 SSD 模型采用多尺度方法,其在不同尺度上的卷积感受野不一样,特别是高层级卷积层,其感

受也很大,因此,对于高层级特征层,其特征提取内容更抽象。而特征提取越抽象,相应的细节信息就越少,从而对小目标的检测不敏感。

卷积层感受野的计算公式如下:

$$S_{RF}(t)=(S_{RF}(t-1)-1)N_S+S_f \tag{9}$$

式中:$S_{RF}(t)$为第 t 层卷积层感受野大小;N_S 为步长;S_f 为滤波器尺寸大小。

另一方面,根据特征图的默认框计算公式可求出某层默认框占输入图像的比例,进而根据两者的映射关系将其映射到输入图像上,这里均选择最小默认框尺寸,以描述不同特征层对输入图像的检测能力。

假如采用 SSD_300×300 的模型,即处理图像的尺寸为 300×300,其特征层主要是 Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2、Conv11_2,则各卷积层感受野大小与各特征层默认框包含的映射图像区域见表 1。

表 1 SSD_300×300 卷积感受野、默认框映射图像区域

Tab.1 Convolution receptive field and the mapping region of default boxes of SSD_300×300

Convolution layer		Feature layer		
Layer	Convolution receptive field	Output scale	Default boxes ratio	Mapping region scale
Conv4_3	92×92	38×38	0.1	30×30
Conv7	260×260	19×19	0.2	60×60
Conv8_2	292×292	10×10	0.38	114×114
Conv9_2	356×356	5×5	0.56	168×168
Conv10_2	485×485	3×3	0.74	222×222
Conv11_2	612×612	1×1	0.92	276×276

从表 1 中可以看出,从 Conv9_2 起,特征层上每一个特征点均由整个图像作为输入产生响应,使得输入图像中目标检测区分性较弱,同时,根据特征层上的默认框映射的图像区域可以看出,Conv9_2 上映射区域已经超过输入图像的一半,即当多个物体同时包含在该区域,则不能被区分,这样的问题对于采用回归的思想做目标检测是不可避免的,同样也出现在 YOLO 中。针对这样的框架,小目标的检测只能通过前面层级的特征图进行识别与定位。因此,SSD 对小目标检测能力较弱。

2.2 多视窗模型建立

为了改善这种现象,在训练得到 SSD_300×300 模型后进行图像的目标检测时,对输入图像进行分区域操作,每一个区域被视为一个视窗,检测时将其放大,模拟人眼近距观测物体时的成像放大机理,以提取更多有效信息,如图 2 所示。

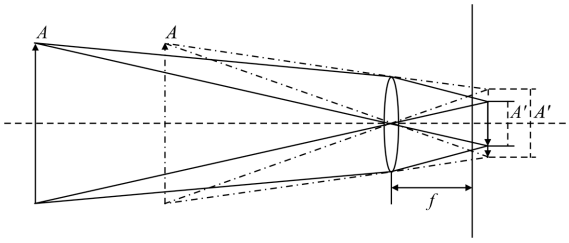

图 2 人眼成像原理

Fig.2 Imaging principle of human eye

图 2 中,物体 **A** 经过眼睛成像之后得到像 **A**′,第一次眼睛观察的过程如图中实线部分所示,当物体 **A** 向眼睛侧移动一定距离之后,第二次眼睛观察过程如图中虚线部分所示,两次观察过程均假设眼睛焦距不变。对比两次观察过程,物体 **A** 的大小没有改变,而经眼睛成像之后的像 **A**′在第二次观察时比第一次观察时大,这种情况下产生视觉响应的视神经细胞会更多被触发,从而使人眼获取的信息增多,细节也会更明显。

为模拟这种原理,对输入图像中感兴趣区域进行分区域放大操作,这里根据位置将输入图像等分为左上、右上、左下、右下四个区域。同时,考虑到中心区域有效信息一般较多,在此基础上加上中部区域,中部区域为与输入图像同中心,并在水平、垂直方向上各截取 1/2 的区域。最后对五个分割区域进行插值,得到与原图一样大的五幅图像,如图 3 所示。

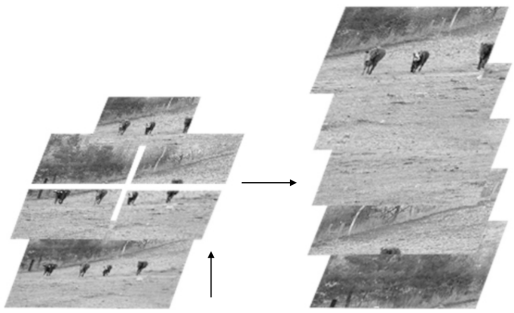

图 3 多视窗切割

Fig.3 Multi-view segmentation

这里将上述模块称为多视窗模块。检测过程中，划分的每一个区域经检测后都得到一组检测结果，最后将得到多组检测目标集。由于原图被切割，必然造成检测结果中存在被切割的现象，需在后端接一个融合模块，负责将每一块区域检测结果融合。

整个检测模型如图 4 所示，其工作流程如下：当一张图片进入模型后，首先经过多视窗模块，将其进行区域切割，被切割成多个区域。进而通过多路检测模型进行分别检测，并将检测结果送入融合模块，得到最终检测结果。

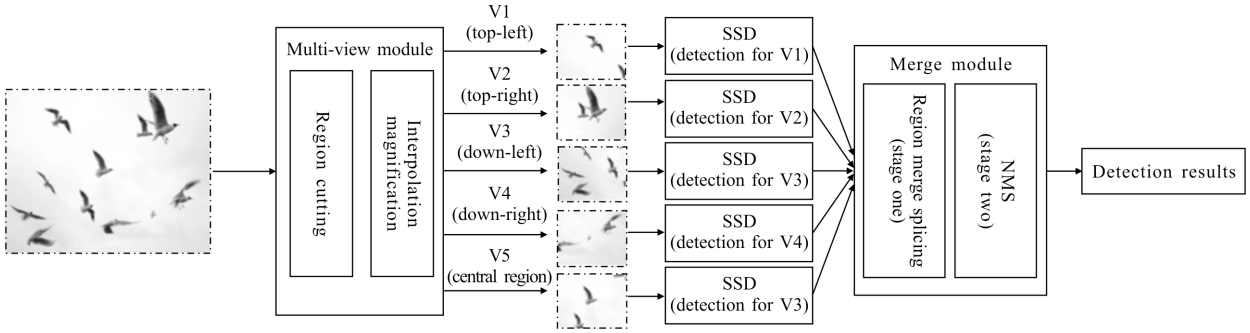

图 4 多视窗 SSD 物体检测模型

Fig.4 Object detection model of multi-view SSD

2.3 模型实现

多视窗物体检测模型主要分为两块：多视窗模块和融合模块。其模型实现主要围绕这两个模块展开。

多视窗模块将原图切割成五个区域，且每一块区域是原图的 1/4，即在检测时，将其采用插值算法等比例放大 4 倍，使得在送入检测模型时检测物体不会发生任何形变，同时，因为采用了插值，提供给检测模块的细节信息相比直接将原图送入检测更多。

融合模块的融合过程主要分为两个阶段。

第一阶段，在切分的四个区域中将根据相邻检测区域中的同类别物体框，是否毗邻水平、垂直轴线，并结合毗邻边界线重合程度判断是否为同一个目标，对于同一个目标需要进行融合拼接。

评价毗邻边界线重合程度，采用边界线重合程度进行判断，其定义如下：

$$L_{overlap}=(L_1\cap L_2)/(L_1\cup L_2) \tag{10}$$

式中：L_1 代表某视窗检测出的 bbox 水平(垂直)轴线附近的边界线；L_2 代表前一视窗毗邻视窗检测出的 bbox 水平(垂直)轴线附近的边界线，如图 5 所示。

如图 5 所示，视窗 V1 与视窗 V2 共同检测出汽车，与视窗 V3 共同检测出人，视窗 V3 与视窗 V4 共同检测出人，其 L_1、L_2、$L_1\cap L_2$、$L_1\cup L_2$ 均在图中进行了标定。计算 $L_{overlap}$，若满足所设定 $L_{overlap}$ 阈值(一般可设置为 10 pixel)，融合被切割在毗邻区域的同一物体，融合之后的物体框种类得分取两者较高分。

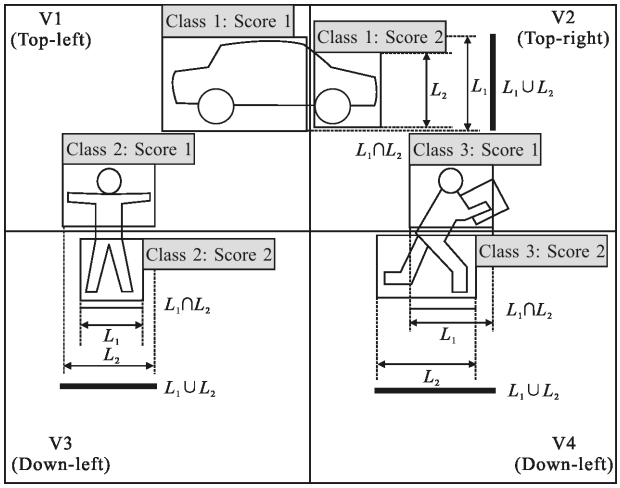

图 5 多视窗毗邻边界线评价

Fig.5 Evaluation of adjacent boundary line in multi-views

具体地，融合拼接过程先做某一方向(如水平轴线)上区域融合拼接，将融合结果加入到检测物体框集合中，并剔除之前进行融合拼接的物体框，然后再在新物体框集合中做另一方向(如垂直轴线)上区域融合拼接。

第二阶段，将第一阶段的融合拼接结果结合中部区域的检测结果共同采用最大值抑制方法(NMS)进行物体框选择。算法的具体操作如下：

Input: Candidate bounding boxes **B** from the stage one
Output: the object bounding boxes $\mathbf{B}_{NMS}$
Initialize the index set **P** of NMS box set, $\mathbf{B}_{NMS}$, overlap threshold $\mathbf{O}_{threshold}$, overlap maximum $\mathbf{O}_{max}$
$\mathbf{P}=\phi$, $\mathbf{B}_{NMS}=\phi$, $\mathbf{O}_{threshold}=t(t<1)$, $\mathbf{O}_{max}=0$
Obtain set of the indexes **I** according to the order of candidate bounding boxes Sorted by the coordinate $\mathbf{y}_2$
While **I** is not null do
  Obtain the last index **I** of **I**: i=**I**[last]
  Set of suppress bounding box **S**=[last]
  Append the last index **I** to **P**: **P**=**P**∪i
  Foreach index **n** in **I** do
    j=**I**[**n**]
    Caculate the overlap **O** (**B** (i), **B** (j)) using the theory of IoU
    If **O**(**B**(i), **B**(j))>$\mathbf{O}_{threshold}$ do
      Append the index **n** to the set of suppress: **S**=**S**∪**n**
      If **O**(**B**(i), **B**(j))>$\mathbf{O}_{max}$ do
        $\mathbf{O}_{max}$=**O**(**B**(i), **B**(j))
        Calculate the area **A**[i] of **B**[i] and the area **A**[j] of **B**(j)
        If **A**[i]≤**A**[j] do
          Remove the last index plast in **P**:**P**=**P**/$\mathbf{P}_{last}$
          Append the index j to **P**:**P**=**P**∪j
Remove the set of suppress in **I**:**I**=**I**/**S**
Foreach **p** in **P** do
  Extract the object bounding boxes $\mathbf{B}_{NMS}$:$\mathbf{B}_{NMS}$=$\mathbf{B}_{NMS}$∪**B**[**p**]

首先初始化 NMS 物体框索引集**P**，覆盖率阈值 $\mathbf{O}_{threshold}$ 和覆盖率最大值 $\mathbf{O}_{max}$。然后将第一阶段得到的候选物体框 **B** 根据坐标 $\mathbf{y}_2$ 进行升序排序，得到排序后候选物体框索引集 **I**，在 **I** 不为空的前提下，循环进行 **I** 中最后一个索引 i(i=**I**[last])指向的物体框 **B**(i)与其他物体框 **B**(j)的覆盖率的计算，在满足覆盖率阈值的条件下添加当前物体框索引至压制索引集 **S**，并判断覆盖率是否大于覆盖率最大值，若满足该条件，覆盖率最大值更新为当前覆盖率，同时，进一步进行 **I** 中最后一个索引指向的物体框与当前物体框的面积的计算，若 **A**(j)大于或等于 **A**(i)，则从 **P** 中移除最后一个索引值，添加索引值j，从 **I** 中移除 **S**，进入下一个循环，直至 **I** 为空，最终得到 NMS 后物体框索引集 **P**，根据 **P** 映射到候选物体框 **B**，得到物体框 $\mathbf{B}_{NMS}$ 作为最后的检测结果。

3 实验与分析

在实际的物体检测中，分类正确、定位准确是所期望的。分类的正确性由预测框的置信度进行衡量，定位的准确性由预测框的坐标信息进行衡量。从物体检测的结果来看，一个优秀的物体检测算法检测出目标应具有尽量高的置信度，同时具备尽量高的准确率和召回率。文中的立足点在于改善小目标的检测，因此，将对比高置信度条件下多视窗 SSD 与 SSD 对小目标的检测性能，以验证文中算法的有效性。这里，选取的置信度阈值为 0.4，即所有检测出的物体的类别置信度均在 0.4 以上。

3.1 VOC2007 数据集小目标测试结果对比

为对比多视窗 SSD 与经典 SSD 的性能，文中从 VOC2007 数据集中共选取 106 张小目标图片，所涉及 ground truth 标注物体 960 个，分别进行经典 SSD 目标检测和多视窗 SSD 目标检测，检测的物体类别 12 种，包含飞机(aero)、鸟(bird)、船(boat)、瓶子(bottle)、汽车(car)、椅子(chair)、牛(cow)、狗(dog)、马(horse)、人(person)、盆栽(potted plant)、羊(sheep)。从中选取了部分具有代表性的检测结果，如图 6 所示。所采用的实验平台：Ubuntu14.04，Quadro K4000 GPU，Xeon E5-2650 CPU。

从图 6 中可以看出，对于小目标的检测，多视窗 SSD 相比于经典 SSD 呈现出以下三个特点：(1) 可以检测出更多的物体；(2) 对于同样识别出的物体，

图 6 经典 SSD(第 1、3 列)与多视窗 SSD(第 2、4 列)检测小目标结果对比

Fig.6 Results comparison of small object detection between the classical SSD(the first column and the third column) and the multi-view SSD(the second column and the forth column)

其置信度更高;(3) 对于相近类别的物体，其准确性更高,修正了经典 SSD 检测时出现的误检测项。而且在检测结果中,有 6 幅图像经典 SSD 没有检测出物体，而多视窗 SSD 均能实现一定数量的物体检测,究其原因,主要是因为这几幅图像里 ground truth 标定物体尺寸较小所致。

下面从定量的角度对多视窗 SSD 与经典 SSD 进行算法性能上的对比，主要以物体检索能力和检测精度两个指标进行评价。

3.2 物体检索能力对比

针对每张图,以目标为主,评价每张图的检索能力,进而求均值,以评估整个系统的检测能力。检索能力一般用 **F** 值(F-measure)表示,其是精度(precision)和召回率(recall)的加权平均。**F** 值的表达式如下:

$$\mathbf{F}=\frac{(\gamma^2+1)\times\mathbf{P}\times\mathbf{R}}{\gamma^2\times(\mathbf{P}+\mathbf{R})} \tag{11}$$

式中:**P** 表示精度;**R** 表示召回率;γ 为权值，一般取 γ=1。

对上述 106 张图的检测结果,绘制多视窗 SSD 与经典 SSD 物体检索能力对比曲线图,如图 7 所示。

从图 7 中可以看出,多视窗 SSD 对每张图像均实现了一定程度的目标检索，且大部分目标检索的结果高于经典 SSD。对 106 张图的检索结果取平均值,求得多视窗 SSD 与经典 SSD 的平均 **F** 值(Average F-measure,AF)分别是 0.729,0.560。其中,经典 SSD 检测时,图像序列 6、32、46、56、71、97 未检测出物体框,图像序列 26、29、50 虽检测出物体框,但是均未满足 IoU 条件(IoU>0.5),使得精度和召回率均等于 0,在这几个位置处,经典 SSD 曲线上的 **F** 值为 0。因此,在小目标检测上,多视窗 SSD 比经典 SSD 的物体检索能力更强。

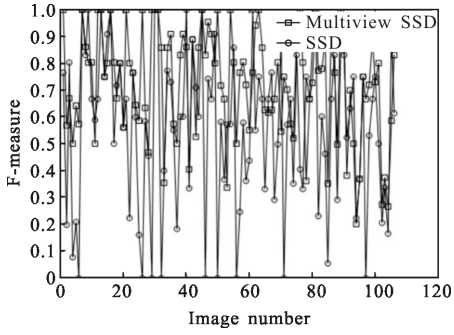

图 7 多视窗 SSD 与经典 SSD 物体检索能力对比

Fig.7 Object retrieval ability comparison between multi-view SSD and classical SSD

3.3 物体检测精度对比

物体检测中，精度的评估一般用 mAP(mean Average Precision)来表示,针对上述 106 张图片,计算经典 SSD 和多视窗 SSD(Multi-view SSD)的mAP,如上所述,文中选取的置信度阈值同样为 0.4,以评估其在高置信度条件下目标的检测精度，结果如表2所示。

从表 2 中可以看出,多视窗 SSD 相比经典 SSD 方法，其检测每一类目标的 Average Precision 基本上都大于后者。同时在无 difficult 目标的检测时,多视窗 SSD 的 mAP 为 0.644,比经典 SSD 检测提高了 0.131;而在包含 difficult 目标的检测时,文中方法的 mAP 为 0.453,比经典 SSD 检测提高了 0.09。因此,在小目标检测上,多视窗 SSD 比经典 SSD 精度更高。

表 2 VOC2007 数据集小目标检测结果

Tab.2 Small object detection of VOC2007 dataset

Method	mAP	Aero	Bird	Boat	Bottle	Car	Chair	Cow	Dog	Horse	Person	Plant	Sheep
SSD	0.513	0.656	0.612	0.620	0.182	0.721	0.447	0.598	0.552	0.286	0.536	0.288	0.661
SSD*	0.363	0.618	0.389	0.296	0.156	0.534	0.340	0.391	0.448	0.111	0.329	0.233	0.506
Multi-view SSD	0.644	0.729	0.662	0.729	0.270	0.875	0.661	0.805	0.662	0.543	0.592	0.452	0.752
Multi-view SSD*	0.453	0.686	0.421	0.348	0.228	0.650	0.502	0.527	0.538	0.211	0.363	0.365	0.599

Comment: Use the "*" to mark the object detection including difficult object or not. The difficult objects are difficult to recognize when experts label the ground truth objects.

综合上述结果，文中方法所提出的多视窗 SSD 在小目标的检索能力和精度上相比经典 SSD 更好。

4 结 论

文中首先阐述了 SSD 的框架及工作原理,然后结合卷积感受野的概念和特征层默认框与输入图像映射关系,分析了 SSD 对小目标检测存在不足的原因,在此基础上,提出了一种改善小目标检测的多视窗 SSD 的算法，经过从 VOC2007 选取的小目标数据集测试，其在物体检索能力与检测精度上均优于经典 SSD 物体检测算法。在物体检索能力上,多视窗 SSD 的平均 **F** 值 (AF) 相比于经典 SSD 提高了 0.169,在检测精度上,不考虑 difficult 目标时,多视窗 SSD 的 mAP 相比于经典 SSD 提高了 0.131,考虑 difficult 时,多视窗 SSD 的 mAP 相比于经典 SSD 提高了 0.09,从而验证了文中方法的有效性。下面将进一步改进其模型,增强其共享机制,改善其时效性,同时在多视窗的条件下结合图像全局信息进一步提高算法的性能。

参考文献：

[1] Erhan D, Szegedy C, Toshev A, et al. Scalable object detection using deep neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2147-2154.

[2] Borji A, Cheng M M, Jiang H, et al. Salient object detection: A benchmark [J]. IEEE Transactions on Image Processing, 2015, 24(12): 5706-5722.

[3] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. Infrared and Laser Engineering, 2017, 46(5): 0502002. (in Chinese)

[4] He Sihua, Yang Shaoqing, Shao Xiaofang, et al. Ship target detection on the sea surface based on natural measure feature of image block [J]. Infrared and Laser Engineering, 2011, 40(9): 1812-1817. (in Chinese)

[5] Merlin P M, Farber D J. A parallel mechanism for detecting curves in pictures [J]. IEEE Transactions on Computers, 1975, 100(1): 96-98.

[6] Singla N. Motion detection based on frame difference method [J]. International Journal of Information & Computation Technology, 2014, 4(15): 1559-1565.

[7] Lee D S. Effective Gaussian mixture learning for video background subtraction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 827-832.

[8] Horn B K P, Schunck B G. Determining optical flow [J]. Artificial Intelligence, 1981, 17(1-3): 185-203.

[9] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//IEE Conference on Computer Vision and Pattern Recognition, 2003, 1: I-511-I-518.

[10] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[12] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]//Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, IEEE, 2005, 1: 886-893.

[13] Panning A, Al-Hamadi A K, Niese R, et al. Facial expression recognition based on Haar-like feature detection [J]. Pattern Recognition & Image Analysis, 2008, 18(3): 447-452.

[14] Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

[15] Zhu J, Zou H, Rosset S, et al. Multi-class adaboost [J]. Statistics and its Interface, 2009, 2(3): 349-360.

[16] Kong T, Yao A, Chen Y, et al. HyperNet: towards accurate region proposal generation and joint object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 845-853.

[17] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[18] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [C]// European Conference on Computer Vision, 2014: 346-361.

[19] Girshick R. Fast r-cnn [C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[20] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [C]// Advances in Neural Information Processing Systems, 2015: 91-99.

[21] Li Y, He K, Sun J. R-fcn: Object detection via region-based fully convolutional networks [C]//Advances in Neural Information Processing Systems, 2016: 379-387.

[22] 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.

[23] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision, 2016: 21-37.

[24] Cai Z, Fan Q, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection [C]// European Conference on Computer Vision, 2016: 354-370.

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//ICLR, 2015.