Voting for Voting in Online Point Cloud Object Detection

<http://www.robots.ox.ac.uk/~mobile/Papers/2015RSS_wang.pdf>

摘要：

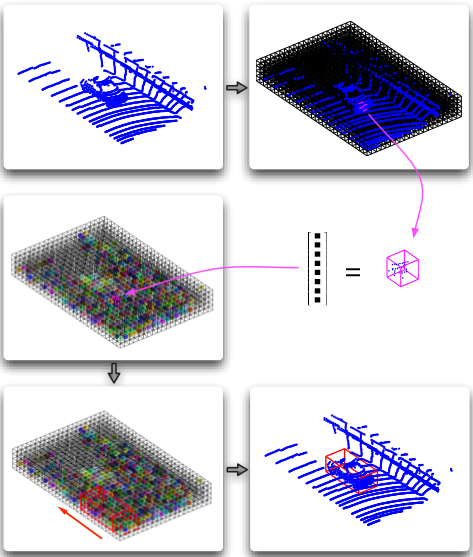
摘要 - 本文提出了一种将计算机视觉中流行的滑动窗口方法应用于三维数据的有效方法。具体而言，通过投票方案利用问题的稀疏性质来在任何方向上搜索所有假定的对象位置。我们证明这个投票方案在数学上相当于稀疏特征网格上的卷积，从而能够以全3D处理任何点云，而不管建立它所需的有利位置的数量。因此，它具有足够的通用性，可以处理来自流行的3D激光扫描仪（如Velodyne）的数据，以及从日益流行的推扫式配置中获取的3D数据。我们的方法是“令人尴尬的可并行化”，并且能够在不到0.5秒的时间内在八个方向上处理包含超过100K点的点云。对于汽车，行人和骑车人的目标分类，所得到的检测器相对于KITTI数据集上的现有技术以及与另一种现有的3D物体检测方法相比实现了最佳的检测和定时性能。

介绍：

目标检测的滑动窗口方法可以说是计算机视觉中最常用的对象检测范例。但是到目前为止，基于激光的物体识别在很大程度上被忽略了。实际上，相同的范例似乎同样适用于3D点云，因为它与2D图像相同。

概念上的区别并不重要，只需要首先将空间分解为三维体素网格，然后像窗口一样通过三维而不是两个窗口来滑动窗口。我们猜想也许有一个贬义因素是额外维度引入的感知计算负担，导致滑动窗口方法的解除在3D中是不切实际的。为了说明，假设我们有兴趣在跨越（100m）2和10m高的区域的点云中检测3D中的对象。如果我们把这个体积分解成（20cm）3个单元，我们将会有1250万个网格单元。一个简单的方法就是在这1250万个单元的每一个单元上放置一个检测窗口的角落，并测试是否限制了一个感兴趣的对象。因此，需要处理约。 1250万个窗口（忽略边界条件，因为这个思想实验是无关紧要的）。即使假设（相当乐观地）单个窗口可以在1μs内处理，这意味着处理单个帧的12.5s的计算负担。使用我们提出的算法，这种情况下的平均计算时间小于0.5秒。

检测过程的说明。 点云（左上）首先被分解成3D网格（右上）。 对于每个被占用的单元格，单元格内的点以及它们的反射值被映射到一个固定的特征向量（右上角给出了一个在右上角和中间左侧高亮显示的被占用单元格的例子）。 未占用的单元通过定义映射到零特征向量。 因此，点云被转换成特征网格（中间左边，每个彩色球体表示为被占用的小区提取的特征向量）。 3D检测窗口随后在所有三维中通过特征网格（左下）滑动，分类器评估每个窗口位置以获取对象的证据。 检测到的对象的点云显示在右下角。 该过程重复每个旋转角度。



实现这种效率的关键在于注意到，与二维图像相比，三维点云的结构存在一个根本的区别：三维点云很稀疏，因为大部分空间都没有被占用。 我们在第四部分展示了滑动窗口检测与线性分类器之间存在一个对偶和一个仅由被占用的单元行使的投票方案，将计算量减少到最小，同时保持精确的数学等价性。 这构成了这项工作的关键贡献，并且导致了我们证明在KITTI数据集上的多个类别上相对于现有技术实现最佳类别检测和定时性能以及与另一个现有3D对象检测做法。

相关工作：

在机器人技术中探索3D数据中物体的检测存在一些工作。

Teichman等人例如，第一段输入激光扫描，然后跟踪在一系列扫描中获得的段。使用助推分类器将每个跟踪的片段序列分类为汽车，行人和骑自行车。

王等人[20]采取了一种相关的方法，其中激光扫描首先被转化为对象部分，并且训练SVM分类器以将该分割成为属于感兴趣对象或背景的前景。然后对前景部分进行聚类，得到对象实例。第二个分类阶段将所得到的集群分类到相同类别的汽车，行人和自行车。 Behley等人[1]也采取分段为基础的方法。首先，从输入点云获得分段的层次结构。然后将这些片段用混合模型进行分类，这些混合模型是通过词袋描述符进行训练的。随后以贪婪方式删除层次结构中的冗余段。赖等。 [11]部署Kinect传感器学习一个稀疏的距离度量标识用于识别预先分段的室内对象与集合套索正则化。

与我们的工作最密切相关的是最近由Song和Xiao [16]提出的滑动窗口方法，他们分享了我们希望通过一个三维像素网格传递三维窗口的愿望。在他们的工作中，通过使用3D积分图像来实现计算易处理性。但是，每个窗口都需要明确测试，以确定是否有更多的被占用的单元比一定的阈值。虽然这比单纯的方法更有效率，但它要求3D网格中的单元总数线性运算。相反，我们的方法在被占用的单元数量上是线性的，空单元被完全绕过。因此我们的方法在计算上更有效率。我们不是第一个注意到滑动窗口检测与线性分类器之间的二元性和投票。莱曼等人[12]使用类似的论点来证明隐式形状模型（ISM）中的投票过程。在他们的框架名为原理隐式形状模型（PRISM）中，有人认为隐式形状模型实际上等价于滑动窗口检测器 - 它们是同一枚硬币的两面。然而，第四部分提出的PRISM框架有三个主要的区别：a）我们的推导中的“投票”没有被投射到一个连续的搜索空间中，它们直接地作为隐藏窗口的独立位置。 b）没有生成码本，特征向量不匹配任何示例。相反，投票只是权重和特征向量之间的标量乘积。 c）最后，也是最重要的，不是概念等价，而是我们在这项工作中展示的是稀疏卷积和投票之间的精确数学等价。

在三维姿态估计[9,7]的单目对象检测工作的主体还通过3D窗口滑动标识的3D边界框来表征对象。然而，不是构建3D特征网格，而是通过将图像平面地投影到物体边界长方体的每个可见面上来实现检测，然后从投影图像中为该面提取2D特征。类似地，3D中另一种常见的物体检测方法是选择或利用单个有利位置，并将来自激光扫描仪的3D数据投影到图像平面上以形成深度图像。然后执行2D滑动窗口操作。作为例子服务[14,15]谁也考虑到外观信息formallike相机images.Laiet等。 [10]通过在RGB和深度数据上的每个视图上的RGB-Ddatabyrunninga2Dslidingwindow检测器的多个视图生成的3D点云中检测对象。然后，通过首先将2D检测分数投影到集成的3D点云并使用体素表示将其融合，来获得3D中的检测结果。相比之下，我们在本文中提出的方法不需要任何投影，只能操作3D数据。因此，我们的方法具有足够的通用性，可以处理来自流行的3D激光扫描仪（如Velodyne）的数据，以及从日益流行的推扫式扫描配置中获得的3D数据，而无需假设有利的位置。

与我们的工作最密切相关的是最近由Song和Xiao [16]提出的滑动窗口方法，他们分享了我们希望通过一个三维像素网格传递三维窗口的愿望。在他们的工作中，通过使用3D积分图像来实现计算易处理性。但是，每个窗口都需要明确测试，以确定是否有更多的被占用的单元比一定的阈值。虽然这比单纯的方法更有效率，但它要求3D网格中的单元总数线性运算。相反，我们的方法在被占用的单元数量上是线性的，空单元被完全绕过。因此我们的方法在计算上更有效率。我们不是第一个注意到滑动窗口检测与线性分类器之间的二元性和投票。莱曼等人[12]使用类似的论点来证明隐式形状模型（ISM）中的投票过程。在他们的框架名为原理隐式形状模型（PRISM）中，有人认为隐式形状模型实际上等价于滑动窗口检测器 - 它们是同一枚硬币的两面。

然而，第四部分提出的PRISM框架有三个主要的区别：a）我们的推导中的“投票”没有被投射到一个连续的搜索空间中，它们直接地作为隐藏窗口的独立位置。 b）没有生成码本，特征向量不匹配任何示例。相反，投票只是权重和特征向量之间的标量乘积。 c）最后，也是最重要的，不是概念等价，而是我们在这项工作中展示的是稀疏卷积和投票之间的精确数学等价。

概述：

我们的3D滑动窗口探测器所需的步骤在概念上类似于基于图像的探测器。图1以一个玩具的例子说明了这个过程 - 在这种情况下，一个真正的3D激光扫描的小部分包含一个感兴趣的对象，一辆汽车。检测输入是以点位置列表表示的3D激光扫描，以及反射率值。首先，将点云转换为如下的特征网格。 3D空间以固定的分辨率离散成网格，每个被占用的单元格被转换成固定的特征向量。未被任何点占用的单元映射到零特征向量。这个定义对于开发问题的稀疏性至关重要。例如，作为一个例子，图1的中间左图显示了在同一图的左上方显示的点云部分上提取的特征网格。在这里，每个彩色球体表示为一个被占用的单元格提取的一个特征向量，不存在球体意味着这个单元格没有被占用，因此它的特征向量是零。注意特征网格的稀疏性 - 彩色球体只占据整个网格的一小部分。

然后在概念上，一个固定大小的3D检测窗口被放置在特征网格的一个角落，并向下滑动所有三个维度。 在每个窗口位置，包含在其边界内的特征向量被堆叠成一个单独的长向量并传递给分类器 （例如，一个SVM）。

分类器然后通过返回检测分数来决定检测窗口的当前位置是否限制感兴趣的对象。 第四节致力于一个关键的数学结果，使这一步易于处理。 与基于图像的探测器相反，尺度在这里不是问题，因为绝对尺度（以米为单位）在3D中是已知的。 但是，轮换是一个问题。 假设感兴趣的物体一般是直立的，也就是说，任何旋转都被限制在垂直轴上，为了能够在任意方向上检测物体，我们将360°分解成N个方向箱并运行相同的检测过程 对于每个定向仓，在旋转的点云上N次。

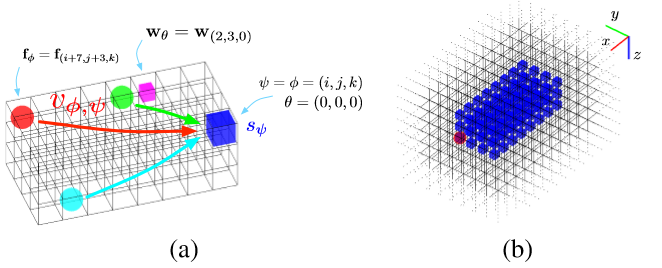
滑动窗口，清晰调整和表决

众所周知，在滑动窗口检测的情况下，线性分类器相当于卷积，因此可以使用快速傅里叶变换（FFT）等标准技术来有效地计算检测分数[3]。 不幸的是，这种技术在我们的3D中并不适用。 空间域的稀疏性并不意味着频域的稀疏性，因此稀疏特征网格的傅里叶变换将是密集的。 我们在本节证明稀疏卷积在数学上等同于投票过程。 这导致了一种计算检测分数的充分方法，充分利用问题的稀疏性给我们带来好处。 我们提供这个作为我们在这项工作中的主要贡献。 请注意，我们要描述的技术只适用于线性分类器的情况。

特征网格自然是四维的 - 每个单元格有一个特征向量，单元格跨越三维网格。我们用flijk来表示单元位置（i，j，k）的第l个特征。有时，我们可以很方便地将在位置（i，j，k）处计算的所有特征统称为向量fijk。为了使演示简单明了，我们用单个变量来引用元组（i，j，k）。 φ=（i，j，k）。如果网格有尺寸（NG x，NG y，NG z），我们可以定义一个集合Φ= [0，NG x）×[0，NG y）×[0，NG z）。这里，符号[m，n]应被理解为在整数集上定义的标准半开区间，即[m，n] = {q∈Z：m≤q<n}，“×”表示设置笛卡尔积。因此Φ是特征网格上的指标集合，并且任何Φ∈Φ索引网格的特定单元格。在这个符号中，fijk可以写成更清晰的形式（这个索引符号如图2（a）所示）。回想一下，如果φ处的单元未被占用，则定义φ= 0。我们可以通过定义一个代表被占用的单元位置的子集的Φ\*Φ来捕获这个概念。因此φ∈Φ\Φ\* =⇒fφ= 0。特征网格是稀疏的。

同样，如果检测窗口的尺寸是（NW x，NW y，NW z），那么定义集合Θ= [0，NW x）×[0，NW y）×[0，NW z） 对于一个检测窗口局部的指标集合，（在图2（a）中也示出了一个例子）。 与特征网格相反，权重是密集的。

1. 说明稀疏卷积和投票之间的二元性。检测窗口的这个位置碰巧只包括三个被占用的单元格（由三个彩色球体表示）。检测窗口的原点（锚点）由拐角处的大蓝色立方体突出显示，恰好与特征网格上的单元位置ψ=φ=（i，j，k）一致。网格位置φ=（i + 7，j + 3，k）处被占用的单元的特征向量如图所示。线性分类器的权重是密集的，而且是四维的。示例位置θ=（2,3,0）的权重向量由一个小洋红色立方体突出显示。所有三个被占用的单元格投票到窗口位置ψ，贡献的分数sψ。我们可以将与窗口中的单元位置相关联的线性分类器的权重表示为wθ。
2. 一个被占领的单元格投票的例子。被占领的单元的位置由红色的球体表示，并且从其接收投票的检测窗口的起源由蓝色立方体表示。这个例子是一个8×4×3的窗口。



为了让我们免于担心边界条件，我们将特征向量和权向量定义为零，如果其索引在边界之外的话。 这将两种情况（特征和权重）中的指数集合扩展到Z3。 除了特征网格Φ上的一组索引之外，我们将需要定义第三组索引Ψ，其包含所有可能接收到非零检测核的检测窗口位置的索引 （0），（0,0,0）（参见图2（a）），那么所有可能的窗口位置的集合可以通过窗口的锚点的特征网格上的所有可能位置的集合来捕获。 因此，所有可能接收到非零检测分数的窗口的锚点的索引的集合由下式给出Ψ= [1-NW x，NG x）×[1-NW y，NG y）×[1-NW z ，NG z）。

在下面的推导中，我们将一致地使用索引φ来索引特征网格上的特征向量，θ来检测窗口中的权值向量，以及ψ来索引窗口位置（网格上的锚点位置）。 我们现在可以得出本节的主要结果。

定理1。 放置在网格位置ψ处的原始窗口的检测窗口可以被写为落入检测窗口内的占用单元格的票数之和。

证明：我们首先根据线性分类器写下检测分数sψ的显式形式。

推论1.三维得分数组s可以写成每个被占用的单元格的投票数组之和。

证明：首先，我们注意到s是一个将Z3中的元素映射到实数（不同窗口位置处的检测分数）的函数，即s：Z3→R。 考虑到这个观点，我们把注意力转回到方程（5），用我们先前对票的定义vφ，ψ=fφ·wφ-ψ，我们到达

现在我们将注意力转移到3D数组vφ的结构上。通过定义，vφ（ψ）=vφ，ψ=fφ·wφ-ψ，这意味着无论何时φ-ψ/∈Θ，vφ（ψ）= 0。回想一下，φ指定投票所源自的被占用单元格的指数，并且ψ投票的窗口位置，这意味着仅在满足φ-ψ∈Θ的位置处的窗口才可能从小区接收到非零投票。现给定一个固定的φ，定义集合Λφ= {ψ∈Z3：φ-ψ∈Θ} = {ψ∈Z3：∃θ∈Θ，ψ=φ-θ}。那么上面的论点限制了单元φ对由Λφ给出的窗口位置子集的投票。参照图（b），图中的红色球体表示被占用的单元的位置，蓝色立方体表示将从φ接收投票的窗口位置。因此包括所有窗口位置，其起点位于与检测窗口具有相同尺寸但是从单元位置向后的窗口中。

评估：

为了便于监督学习，我们利用公开的KITTI数据集[8]。 KITTI数据集的物体检测基准提供同步摄像机和Velodyne帧，同时在图像和激光数据中注明对象。 我们感兴趣的特别之处在于激光数据中的注释是作为完全定向的三维边界框给出的，以规范的方向界定感兴趣的对象。 我们评估所提出的三维滑动窗口检测器在汽车，行人和自行车等级上的性能，并且基于通用评估指标在其绝对性能方面表现出优越的性能，并将相对性能与现有的基于分段的方法 到3D对象检测[20]，并在KITTI对象基准测试网站上发布结果。

1. 训练

标准的KITTI物体检测基准包含一个标记的训练集和一个标记的测试集。 但是，为了评估目的，测试装置上的标签被阻止。 由于我们感兴趣的是对滑动窗口检测器在3D数据上的性能的公平评估，而KITTI主要是一个视觉数据集，所以我们从KITTI中公开提供的标记数据创建我们自己的训练和测试数据集 即原来的“训练”数据集），通过随机将其分成两部分，然后根据更适合评估三维检测的度量测试检测器的性能（参见第VII-B节）。

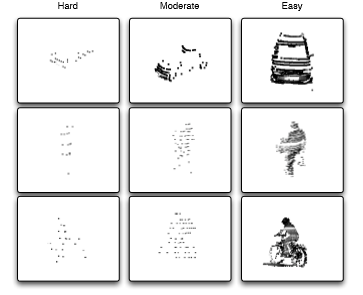
具体来说，我们随机将可用的7481个标记帧分成80/20比例的训练和测试。表1列出了所得训练集和测试集中包含的帧数以及其他信息。对于线性分类器，我们部署一个线性SVM，并使用LIBLINEAR库[5]进行训练。从训练数据中随机抽取最初的一组否定实例（等于正实例的数目），注意不要与任何正实例重叠。以这个初始训练集为例，我们采用基于图像的物体检测器的标准硬阴图挖掘技术（例如[17,2,6]）。具体而言，分类器首先在初始训练集上进行训练。训练结束后，将分类器应用于所有训练框架。所有训练帧上的这个分类器的所有假阳性检测被整理，并按检测分数的降序排序。第一个N（或所有的误报，如果有少于N），然后采取并添加到负面的例子集合。然后用这个更新的训练集重新训练分类器，并且这个过程可以迭代预定义的循环次数。在我们所有的实验中，我们把N定义为10000，进行20轮硬性负面开采。

滑动窗口方法的一个缺点是可能在离散过程中引入人为因素。 由于窗口位置只能在离散的特征网格上进行搜索（并且不能对其进行检索），因此不可能在检测窗口中正确捕获一个对象。 然而，训练的正面例子是从人工标签中提取出来的，因此所包含的对象是居中和朝前的。 为了弥补这种差异，对于每一个正面的例子，我们随机抽取10个稍微翻译和旋转（关于垂直轴）的版本，并将它们附加到用于训练的一组正面例子。

1. 评估策略

KITTI数据集对三维激光数据提供的对象标签是全面的，因为除了明显的对象实例之外，由于离激光距离较远传感器也包括在内。所包括的物体有时可能像仅通过少量激光测量所描述的那样具有挑战性（参见例如图3的左栏）。这激励我们将标记的对象实例划分为与原始KITTI规范[8]类似的不同难度级别，同时尊重数据集中的完整标签集合，而不是对检测系统提出不合理的要求。最初的KITTI规格是专门为基于视觉的检测系统量身定制的。在这里，我们首先仔细研究3D激光数据中提供的标记对象实例类型的数据集，并据此设计合适的准则，将对象分为简单，中等和难度级别。

来自不同难度训练集的标记对象实例的例子。 左列：硬（但不是中等），包含测量数量<m <50的实例。中间列：中等（但不容易），包含测量数量的实例50≤m <150.右列：简单，包含测量数量的实例 m≥150。详见文字



1. 检测性能

提出的滑动窗口检测器是按照章节VII-A中概述的程序用训练集进行训练，并在测试集上进行评估。检测器只有三个参数，格网分辨率δ（网格单元的边长），角度单元的数目N和非最大抑制的重叠阈值（参见第VI节）。在我们所有的实验中，我们设定δ= 0.2m，N = 8，对于汽车类别= 0.01，对于行人类别设定为= 0.5，对于自行车类别设定为= 0.1。依赖于类的参数被选择来反映该类别的对象之间的期望接近度。图4给出了20轮硬核负面挖掘训练后的三个目标类别的结果，每个类别在测试集的三个不同难度级别上进行评估。图中显示的精确回忆曲线是通过改变检测阈值σ产生的（参见第VI节）。正如人们所期望的那样，由于所有对象类别的评估难度都有所降低，因此检测器的性能会更好，在易于使用的情况下，性能最佳。

F.与在KITTI数据集上评估的其他仅激光方法的比较

正如前面提到的，由于缺乏已经建立的三维物体检测基准，目前很难将所提出的方法与其他基于激光的检测方法进行基准比较，因此基于视觉的检测器的标准基准KITTI数据集。为了扩大我们的比较研究的范围，我们已经将所提出的3D检测器的结果提交给2D KITTI对象基准，通过将所获得的3D检测投影到具有所提供的校准参数的图像平面1。我们在这里必须强调，尽管在相同的测试数据集上按照相同的程序进行评估，但与其他公开的基于视觉的方法的结果的任何比较都是不可避免的严重偏倚。我们提交的作品没有利用任何图像数据，而且评估标准是根据原始的关于视觉的难易程度来定义的。由于传感器模式的固有差异，试图在完全公平的基础上比较基于激光的和基于视觉的方法即使不是不可能也是具有挑战性的。例如，在视觉上困难的外观在激光中可能并不困难，反之亦然。

然而，目前在基准测试网站上存在两种（也是仅有的）其他激光测量方法，以类似的方式公布其检测结果。 表三是KITTI网站公布结果的摘录，将我们提交的结果（Vote3D）与这些方法直接比较。 对于每个对象类别，表格列出了基准中方法的等级，以及中等，简单和困难难度水平的平均精度（提交按照KITTI基准评估的中等难度的平均精度排序， 因此排序）。 CSoR提交是匿名提交的，仅在汽车类别上进行评估。 每列中粗体的数字表示该列的顶部条目（更高的等级或更高的平均精确度）。 从表中可以看出，所提出的滑动窗口检测器在所有列上胜过胜利。