我们距离解决行人检测到底有多远？

摘要：

受到最近行人检测发展的启发，我们研究了目前最先进的算法和“perfect single frame detector”之间的差距。我们通过为行人检测建立一个human baseline 来进行我们的分析。（加州理工学院数据集），并人工的将顶级检测器反复发生的误差聚成一类。我们的结果描述了定位和背景vs前景的误差。

为了找出定位误差我们研究了训练标注噪声对检测器的影响，结果表明我们还可以在已经清洗过的训练数据上做一小部分提高。为了找出背景/前景的区别，我们研究了行人检测的卷积网络，并且讨论了哪些因素影响了检测器的性能。

除了深入的分析，我们还报告了该数据集上最好的表现并且提供了一个新的处理过的训练和测试标注。

1. 简介

物体检测这几年一直备受关注。行人检测就是一个典型的子问题，它依然是研究中的热门话题出于其广泛的应用。

尽管关于行人检测的研究已经很多了，最近的论文依然展现出了重大的提升，这也表明了该研究还未达到饱和点。本文中我们分析了最先进的算法和最新的human baseline 之间的差距。结果表明目前检测器与人类的判断差距十倍之多。我们专注于研究哪些因素可以帮助减少这个差距。

我们分析了哪些顶级行人检测器的失败案例并探究为了改善它的表现需要哪些改变。我们展示几个不同的分析，包括人类检测，自动的问题案例分析和oracle 的实验。结果表明定位是结果高误报率的主要原因。我们通过手工清理训练标注和一些对剩余训练样本的算法来提高训练集的对准质量以处理这一方面的问题。

对于背景vs前景辨别的问题，我们研究了行人检测的卷积网络，并讨论了哪些因素影响了检测器的性能。

* 1. 相关工作

最近几年，很多人在提高行人检测器的性能上做出了不少努力。随着ICF检测器的成功，不少变种都被推荐并展现了重大的改进。从最近的一个关于行人检测的综述中我们可以得出结论，改进的功能一直在推动着检测器的性能并很有可能会持续这样下去。这也同样表明了光流和环境信息是对图像特征的补充并在未来可以提高检测的准确率。

通过对模型每次在外部数据上训练的微调，卷积网络也达到了最好的表现。

最近大部分的论文关注于新奇和更佳的结果，但忽视了对系统的分析。一些分析工作可以在一般的物体检测上找到；相反的，在行人检测的领域，这类型的分析很少有人做。在2008年，一个在相对较小的INRIA数据集上的失败分析被提供出来。被认为是2012年最好的方法在Caltech数据集上的测量比目前我们研究的方法有10倍的误报率在20%的查全率上，并且没有方法可以达到95%标准。

由于行人检测在近几年中已经被改进了不少，一个基于最先进检测器的更深层次、更综合的分析就变得尤为重要，以便指引未来的研究方向。

* 1. 所做出的贡献

我们主要做出的贡献如下：

1. 我们提供了一个对最先进的行人检测系统的详细分析，包括对失败案例的深入了解。
2. 我们给Calech 行人基准提供了一个human baseline；还包括一个清理过的标注来作为新的、高质量基础标准来作为训练集和测试集的基准。数据也将会被公开。
3. 我们分析了训练数据的质量对检测器的影响有多大。更特别的是我们对对齐程度和误标注的多少可以提高系统性能进行了量化。
4. 使用对分析的见解，我们探索了变现最好的一些方法的变种：filtered channel feature detector 和 R-CNN detector，并展示了对基线的提高。
5. 准备工作

在探究我们的分析之前，让我们先表述一下所使用数据集、度量以及我们的基线检测器。

* 1. Caltech-USA 行人检测基准

在已有的行人数据集中，KITTI和Caltech-USA是目前最出名的。本次研究中我们使用了由一辆穿过美国洛杉矶街道所记录下的2.5小时30HZ视频记录组成的Caltech-USA 标准。该视频的标注达到了总数为350000个边界框、覆盖了2300名行人。检测方法在一个包含4024帧的训练集上被评估。所提供的评估工具箱根据标注的大小，遮挡的程度和宽高比来生成不同测试子集的图像。为训练所建立的程序使用视频每三十帧，总共4250帧，包含1600个行人的剪切图作为数据。最近，那些可以生成更多训练数据的方法已经被使用来更加精细的对视频进行抽样，比标准的“1x”设置可以产生10倍的数据。

MRO,MRN 在标准的Caltech 评估中，缺失率平均在[10^-2，10^0]FPPI（False Positive Per Image）左右。这个度量并没有很好的反映出在定位上的误差（最低的FPPI范围）。为了得到一个更加完整的评估，我们把评估FPPI范围从传统的[10^-2,10^0]扩展到[10^-4,10^0]，并用符号MRO-2和MRO-4来表示。O代表了“原始标注”。在3.3节我们将会介绍新的标注并且标记完成的评估为MRN-2和MRN-4。我们希望MR-4标准可以在检测器强大时变得更加重要。

* 1. Filtered channel features detector

在本文的分析中，我们考虑到了所有目前为止Caltech Pedestrian 标准中提供的方法，一直到最近一次大会（CVPR2015）。如图一所示，目前最好的方法是Checkerboards，并且大部分表现出色的方法都是它的同族。

Checkerboards 检测器是ICF的一般化展现，它在把数据带入提升决策树之前过滤了HOG+LUV特征通道。

表1中我们比较了几个与ICF同族的检测器我们可以看到一个很大的提升那就是通过引入在特征通道中引入过滤器以及优化过滤器库可以把MRO-2从44.2%减少到18.5%。

最近表现较好的卷积网络方法对于基础的检测方案都很敏感，因此我们首先专注于优化filtered channel feature detectors。

Rotated filters 在涉及训练新模型的实验中，我们根据LDCF的代码基础使用了自己对Checkerboards的实现。为了降低训练时间我们减少了过滤器的数量，从原始的61个降低到了6个。我们所称的RotatedFilters是一个LDCF的简化版本，应用于三个不同的规模。更多关于过滤器的详细材料将会在补充材料中给出。如表1所示，RotatedFilters比LDCF要好的多，只比Checkerboards相差一个百分点，并且在训练和测试集上的运行时间要快六倍。

其他原因 这篇review展示了环境和光流信息可以帮助提高检测。但是，随着检测器品质的提高，返回的结果就会受到这些原因的腐蚀。如果不对这些原因进行工程再造，检测所获得的就必须来源于核心的检测器。

1. 分析最先进的方法

在这一节中我们将评估目前剩余可用发展的更低界限，分析目前行人检测器的误差并且提出新的标注以度量未来的发展。

* 1. 我们到达饱和了吗？

行人检测的发展这几年并没有减缓的迹象，尽管最近有一些骄人的成果。就当前的水准来说还有多少发展值得被期待呢？要回答这个问题，我们提出了使用human baseline 作为下界。我们请一些领域专家来在Caltech-USA的测试集上手工的“检测”行人；机器检测算法应当可以至少达到人类的标准，最终超越人类。

Human baseline 协议 为了确保对当前的检测器做一个公平的比较，我们专注于每一帧单目检测的设置。每一帧被随机的提供给标注者们，并且不能从原视频中获取周围帧的信息。标注者们必须根据行人的外形和单帧的环境信息而不是移动的线索。

The Caltech benchmark 对检测框的长宽比做了规范化。因此我们人类标注是通过画从头到两脚之间的点的线来完成的。一个中心点与手绘轴线中心点一致的边界框就会自动的生成。如图2.这个过程保证了边框可以很好的包围住物体。

为了检测两个标注者的一致性，我们为一个测试图片的子集复制多个标注，然后将这些分开评估。IoU>=0.5的结果将会被统一成一个单一的边界框。

结论 在图3中，我们在不同的测试数据子集上比较了我们的human baseline和另一个较好的方法，我们发现在所有设置下，human baseline都要比最先进的检测器表现要好，这也表明了自动化方法依然有提升的空间。

* 1. 失败分析

既然对于目前的检测器来说都有提升的空间，你可能会问：什么时候他们会失败？在这一节我们将分析在大部分测试集子集上表现非常好的Checkerboards的检测误差。由于图1中大部分顶尖的方法都与ICF同族，我们也认为他们会有相似的表现。那些使用基于ICF检测器方案卷积网络的方法同样也会被影响。

* + 1. 误差来源

检测器主要有两种类型的误差：误报和漏报。在这些分析中，我们调查了在0.1FPPI下的误报和漏报，并手工的把他们分到不同的组中。一共有402个误报和148个漏报结果根据错误类型被分类。

误报 再检查过后，我们最终把误报聚成了11类，如图4a。这些类归于三组：定位误差、背景误差和标注误差。

背景误差是最普遍的，主要是一些垂直的建筑物（如图5b），树叶和交通灯。这表明了检测器需要在更好的垂直环境上改进，提供对大型建筑物的可视化以及粗略的高度评估。

定位误差主要是双重检测（高分数的检测结果覆盖同一行人，如图5a）。这表明了改进的检测器需要更多局部的反馈(峰值分数图)和不同的非极大值抑制策略。在3.3节和4.1节中我们将探究怎么提高检测器的定位。

标注误差主要是没有忽略地区，还有一部分缺失行人标注。在3.3节我们将再次回顾Caltech标注。

漏报 如图4，我们的聚类结果展示了检测小型和遮挡物体的难题。我们假定侧面的行人和骑自行车的人的低得分由于数据集的偏差。这些情况在训练集中表现不足（大部分不是骑自行车的，而是走在人行道上，与车平行）。用额外的图片来在训练集上补充这一方面的信息应该是一个有效的策略。

为了更好的理解较小的行人问题，我们测量了每个检测物的大小，模糊程度和对比度。我们发现较小的行人都普遍饱和的（暴露过度或不足）和模糊的，因此假设这可能是一个弱检测的潜在因素（除了那些只有很少像素就得出结果的）。然而我们的结果表明，情况并非如此。如图4c，似乎较低的检测分数和低对比对并没有相关性。这对于模型的情况也同样适用，更详细的绘图会在补充材料中给出。我们得出结论：像素数量过少才是检测困难的真正来源。因此提高小型物体的检测需要正确利用像素信息，不管是窗口中的还是周围环境的，又或是跨越的时间。

结论 我们的分析表明误报的原因十分明确，可以针对不同情况使用上述的建议策略。一小部分的漏报同样可以找出原因，即使小而封闭的行人依然是最主要的问题。

* + 1. Oracle test cases

在3.2.1节中的分析主要是在误差的原因。对于area-under-the-curve 这类指标，例如Caltech中使用的案例，高分误差比低分误差更加重要。在这一节，我们使用oracle test cases来直接度量定位误差和背景vs前景误差对于检测质量度量的影响。

在oracle case中的定位问题，所有的与ground truth 有所交叠的误报在评估中都被忽略了。在oracle case中的背景vs前景问题，所有没有和ground truth交叠的误报都被忽略了。

图6a展示了修正定位误差在低FPPI区域的提高；修正背景误差在高FPPI区域对结果有提高。两种类型都修正则会产生零误差，尽管由于double log plot，这不会被立刻看到。

在图6b中我们通过修正定位或背景问题可以获得在MR0-4上的增益。在将八个表现最好的方法进行比较时，我们发现大部分方法都会在修正其问题后得到重大的提升。需要注意的是由于log-log的数字性质，定位和背景deltas的综合加起来不到缺失率。

结论 对于大部分表现良好的方法来说，定位和背景vs前景误差对于检测质量的影响一致。他们同等重要。

* 1. Improved Caltech-USA annotations

当我们在IoU>=0.8的条件下评估我们的human baseline时，我们发现了如图三的表现下降。原始的标注协议是基于多个帧插入的稀疏标注，并且这些稀疏的标注不一定位于要评估的帧上。在仔细的检查过后我们发现这种的插值会产生一种在标注上的系统偏移。

人类走路会伴随着自然的上下振动，并不是使用线性插值所建的模型，因此在大部分帧中都转移了边界框标注。这个影响在IoU>=0.5时是发现不了的，但这些标注中的噪声都会是提高物体定位时的障碍。

在3.2.1中定位的问题和标注的误差使我们想要为Caltech行人数据集创造新的标注集。我们的目的有两方面；一方面我们想提供一个对于最先进方法更加准确的评估，特别是对于问题“最后20%”合适的评估。另一方面，我们想要评估对于标注的提高量到达多少可以使检测结果更佳。我们将在4.1节评估这两个方面。

新的标注协议 我们的human baseline 专注于使用单帧的方法来进行公平的比较。我们的新标注在测试集和训练集上都完成了，并关注于高的质量。标注者可以观看整个录像来确定某个行人是否出现，他们被要求在覆盖人群的区域标记出忽略的区域，是人型但不是人（海报，雕塑等）的区域，以及那些不能确定是否包含行人的区域。每个行人的标注都是通过从头部到两脚之间在中点画一条线来完成的，和human baseline一样。The annotators must hallucinate head and feet if these are not visible.当一个行人不是完全可见时，他们同样必须用矩形标注出最大可见区域。这就可以与原始的Caltech 标注使用同样的方式来估计遮挡的水平。新的标注与human baseline的一些边界框相同（当不需要修整时），因此human baseline不能被用来在不同IoU阈值的新测试集上做分析。

总的来说，我们新的标注与human baseline主要有以下几方面不同：训练集和测试集都被标注了，忽略的区域和遮挡的区域也被标注了，整个录像的数据都被用来决策，并允许同一个图像的多个修订。

在做完整个独立的标注集后，我们通过与旧标注的交叉验证来整理出新的标注。任何没有被算在新集合中的正确的旧标注也被加了进来。

我们新的标注改进了已有标注的几个类型的误差，比如不对准，漏标注，误标注以及对“忽略”区域的不一致使用。我们新的标注将会公开。附加“原始vs新标注”的例子将在补充材料中给出，以及可以对他们逐帧查看的可视化软件。

更好的对准 在表3中展示了我们的新标注比原始标注定位更加准确的定量证明。我们通过正确检测和已知的标注集之间的中间IoU值总结出检测器的对准质量。当使用原始标注进行评估时，只有用原始标注训练的模型才有较好的定位。但是，当使用新标注进行评估时，通过INRIA数据训练的模型和新标注训练的模型都达到了很高的定位准确度。这也表明了我们的新标注确实对准水平更高，就像INRIA标注优于Caltech一样。

更佳详细的多检测器IoU曲线将会在补充材料中给出。在4.1节中将会描述RotatedFilters-New10 x entry.

1. 对最好方法的改进

在这一节中我们利用分析的见解来提高我们基线检测器的定位和背景vs前景识别。

* 1. 训练标注的影响

我们想利用手上的新标注来了解标注质量对检测质量的影响。我们将使用不同的训练集训练ACF和RotatedFilters模型并在原始和新的标注上评价。由于两种检测器都是通过提升训练的因此本身对标注噪声敏感。

剪枝的好处 表4展示了使用原始的，新的和修剪过的标注进行训练的结果。正如期望的，在原始/新标注集上训练并在原始/新标注集上测试的模型比训练和测试在不同标注集上的模型表现要好。为了更好的理解新标注带来了什么好处，我们构建了一个混合的标注集。修剪过的标注集是一个可以解除移除误差和提高对准的中间点。修剪过的标注集是通过匹配新的和原始的标注（IoU>=0.5），把新标注集中没有的原始标注都标记为忽略区域，并且把原始标注中没有的新标注加上而生成的。

从原始到修剪的标注，主要的改变就是移除标注误差，从修剪的到新的标注，主要的改变是更好的对准。从表4中可以看出ACF和RotatedFilters都从移除标注误差中收益，甚至是MR0-2.这表明了我们的新训练集比原始训练集处理的更好。我们从MRN-2可以看出更强的检测器从更好的数据中收益更多，并且检测质量最大的增益来源于移除标注的误差。

对准的受益 ICF族的检测器都从增长的训练集中受益，使用十倍的数据比使用一倍的数据要好。为了剩余的9倍数据使用新的1倍新的标注，我们用新的标注集训练了一个模型，并用这个模型对原始的9x部分标注进行重对准。因为新的标注有着更好的对准，我们希望这个模型可以弥补在原始标注上些微的位置和范围误差。详细内容在补充材料中。表5展示了使用自动对准处理以及一些递减的案例的结果：使用原始10x的数据，使用在原始10x数据上训练的模型对齐进行自对准，以及使用新标注的一部分对原始10x进行对准。结果表明使用检测器模型来对全部数据对准进行改进确实是有效的，并且对准程度更好的训练数据会产生更好的检测质量。这也与3.2节中分析的一致。我们已经在使用一个在二分之一新标注数据上训练的模型，这一模型比使用原始标注的模型更加强劲。

我们把使用新标注和对准后的9x数据训练的RotatedFilters模型命名为RotatedFilters-New10 x。在表3中，这个模型也达到了高中位数的true positives IoU，表明了该模型确实在测试中更为准确。

结论 使用更高质量的标注来训练模型可以提高全体的检测质量，归功于改善的对准程度以及标注误差的减少。

* 1. 行人检测的卷积网络

3.2节中的结果指明通过专注于核心背景vs前景识别的任务还有提升的空间。近期的工作显示了卷积网络在行人检测中的竞争性。我们也对卷积网络进行了分析，并且探究了是什么可以改善检测的质量。

AlexNet and VGG 我们考虑了两种卷积网络。1）AlexNet 2)VGG16模型。两者都在ImageNet上提前训练好并在Caltech 10 x上使用SquaresChnFtrs进行过调参。两者都基于开源网络并都属于R-CNN架构。虽然他们的训练/测试时间架构有一点不同，我们期望结果的不同主要是由于他们各自的识别能力不同。

表6展示了我们改善了检测标准，AlexNet不能提供一个持续的增益，最总比ICF检测器的结果更糟。相似的VGG对较弱的标准提供了更大的增益，但是随着标准提高，来自卷积网络重得分的增益最终停滞。

在细致研究了结果曲线，我们发现AlexNet和VGG都会使背景实例分数较低，同时产生大量的高得分误报。ICF检测器可以提供高查全proposals，以及物体周围的误报具有低的得分，但是卷积网络在给予窗口周围的正确检测低分方面有困难。换句话说，尽管他们微调过，卷积网络分数图比proposal的更模糊。我们假设这是AlexNet和VGG结构上固有的限制，由于其内部的特征池化。从卷积网络中获得“峰值”相应很有可能需求使用不同的架构，可能与需要准确像素输出的语义标注和边界估计任务使用的相似。

幸运的是，我们可以使用边界框回归来弥补卷积网络评分中空间分辨率的不足。在VGG中加入边界框回归，并应用第二轮非极大值抑制，有“contracting the score maps”的效果。相邻的那些在生成多个强误报的方案，现在也瓦解成一个单一的高得分检测。我们对第二个NMS使用了通常的IoU>=0.5的合并标准。

表6中最后一列展示了边界框回归+NMS在提供额外输入建议上的增益是有效的，即使是对于我们最好的检测器RotatedFilters-New10x.在原始标注上RotatedFilters-New10x+VGG达到了14.7% MR0-2，这已经提升了很多。

图9重复了我们3.2.2中对oracle tests的卷积网络结果。可以看出VGG极大的减少了背景的误差，同时稍微提高了定位的误差。

结论 尽管卷积网络在图像分类和一般物体检测中有着很好的结果，他们似乎在对小物体定位检测时有限制。边界框回归（和NMS）是使用目前的结构回避这种限制的重要组成部分。即使使用过强卷积网络后，背景vs前景依然是主要的误差来源；表明了在神经网络的原始分类能力上还有提高的空间。

1. 总结

本文中，我们努力去分析顶尖的探测器在Caltech数据集上的误差。我们通过human baseline 对期望提升空间的最低标准做了一个量化。还有10x 的距离。为了更好的度量下一步的检测发展，我们提供了一组处理过的Caltech训练和测试标注集。

我们对顶尖方法的失败分析表明他们大部分的误差都有很好的表征。误差的特点也可推出如何制作更好的检测器的特定的建议。（在3.2节中曾提到；如增加行人侧视图的数据，或在垂直方向拓宽检测器的接收区域）。

我们通过度量更好的标注对定位准确率的影响以及通过探究使用卷积网络对背景vs前景识别的提高来特别的提出一些问题。我们的结果表明适当的训练ICF检测器可以使其在对准程度上得到提高，对于行人检测，卷积网络与定位的问题可以通过边界框回归来解决一部分。无论是原始还是新的标注集上，描述的检测方法达到了很好的表现，见表7。

我们希望本研究中提供的见解和数据可以减少机器与人类在行人检测任务上的差距。