**行人检测：一种对最好方法的评估**

**摘要：**行人检测是计算机视觉中的核心问题，关于其的一些应用对生活的质量有潜在的积极影响。近几年来，单目图像行人检测方法的数量在稳步增长。但是，使用的多种多样的数据集和广泛变化的评估协议，使得直接的比较十分困难。为了处理这些不足，我们在一个统一的框架上提出来一种对最好方法广泛的评估方法。我们做了主要三方面的贡献：（1）我们组合了一个大型，有良好标注的，现实的单目行人检测数据集并对城市场景下行人大小、位置和遮挡模式的统计进行了研究。（2）我们提出了一个精确的每帧评估方法，该方法允许我们进行探测和信息比较，包括度量关于尺度和遮挡的表现。（3）我们评估了16个横跨6个数据集预先训练的最好的检测器的性能。我们的研究使我们可以评估最好的方法并且为衡量未来工作提供了一个框架。我们的实验表明尽管现在已有众多发展，检测器性能依然还有很大的提升空间。特别的，检测器在低分辨率和行人部分遮挡时表现不佳。

**索引词：**行人检测，物体检测，基准，评估，数据集，Caltech 行人数据集

1. **简介**

人是机器环境中最重要的组成部分，赋予机器人机交互能力是现代工程最有趣和潜在有用的挑战之一。因此，检测与追踪行人是研究中重要的部分，机器视觉必定起到关键作用。主要应用包括机器人学，娱乐，监测，关心老人与残疾人以及基于内容的索引。仅在美国，每年35000起交通事故中，有5000起涉及行人，因此对为行人监测建立自动化视觉系统有着巨大兴趣。

虽然有很多为行人检测的机器视觉方法的研究从未间断，广泛的评估协议以及不同数据集的使用使得直接的比较十分困难。最基础的问题例如“目前的检测器工作的好吗？”，“最好的方法是什么？”，“主要失败的模式是什么？”以及“最有成效的研究方向是什么？”都不是很好回答。

我们的研究目的在于解决这些问题。我们专注于在单个单目图像上进行行人检测的方法；对于检测器如何并入完整系统的概述，读者可参考[2]。我们的方法是三方面的：我们对一个从城市交通机器动车导航中收集的大型行人图像数据集进行收集、标注以及研究；我们开发了信息评估方法并且指出之前实验中的缺陷；最后，我们在6个公开数据集上比较了16个预先训练好的行人检测器，包括我们自己的。我们的研究使得我们可以对最好的方法进行评估以及为未来研究提出了方向。

本研究的所有结果以及对其进行重生产的数据和工具都上传到了如下网站：[www.vision.caltech.edu/Imge\_Datasets/CaltechPedestrians/](http://www.vision.caltech.edu/Imge_Datasets/CaltechPedestrians/).

* 1. **贡献**

**数据集：**在早期的工作，我们介绍了Caltech行人数据集，该数据集包括了35000个在25000帧上标注的行人边界框，迄今为止依然是最大的数据集。遮挡和对应时间也被标注了。利用大量的ground truth，我们分析了行人尺寸，遮挡和位置的统计信息并且帮助建立了检测系统必须运行的条件。

**评估方法：**我们的目标在于以现实和公正的方式来量化以及评估检测器性能。

**评估：**我们评估了16个有代表性的最好的行人检测器（之前我们评估了7个）。我们的目标是选择多个在最初报告的性能方面有前途的检测器。我们避免重训练或调整检测器，确保每个方法都是其作者调优过的。除了整体的性能，我们还探究了在不同级别的尺度和遮挡、清晰可见的行人下的检测率。更进一步的，我们测量了定位准确度以及分析运行时间。

为了增加我们分析的范围，我们也使用了一个在6个附加行人检测数集上（包括ETH,TUD-Brussels，Daimler，INRIA和两个Caltech 数据集见图1）统一的评估框架来对16个检测器进行基准测试。通过在多个数据集上进行评估，我们可以评估检测器性能以及对结果统计信息的重要性进行分析，更一般的，就检测器和数据集本身而言得出结论。

两组团队最近发布了一些调查作为我们研究的补充。Geronimo等在[2]中对高级驾驶辅助系统，重点是全系统的行人检测进行了全面的调查。Enzweiler和Gavrila公布了Daimler数据集以及三个检测器的附带评估，并进行了将检测器集成到全系统的附加实验。我们反而专注于对最好检测器的一个更彻底、更详细的评估。

本文组织如下：我们介绍Caltech行人数据集并在第二节中分析其统计数据；2.4节中将会给出现有数据集的一个比较。第三节中我们详细讨论了评估方法一个关于行人检测器的调查将会在4.1节中给出，在4.2节我们使用我们的评估方法对代表性的16个最好检测器进行讨论。第五节中我们报告了性能评估的结果，包括在不同条件下使用Caltech数据集以及在6个附加数据集上。我们在第六节中通过对行人检测中最好方法的讨论来进行总结。

1. **Caltech行人数据集**

有挑战的数据集是计算机视觉发展的催化剂。Barron[8]等和Middlebury[9]光流数据集，Berkeley Segmentation 数据集，Middlebury Stereo数据集以及Caltech101，Caltech256和PASCAL物体识别数据集都提高了性能评估，增加了挑战，并有助于推动他们在各自领域的创新。同样的，我们介绍Caltech 行人数据集的目的在于提供一个更好的记住以及帮助确定目前检测器失败的条件并依此将研究重点放在这些困难的情况上。

* 1. 数据收集和Ground Truthing

我们从在城市环境下一辆行驶通过正常交通的机动车拍摄的30HZ录像中收集了大约10小时数据。（图2（a）展示了摄像机设置）CCD视频分辨率为640x480，不出意料的，整体图像质量比可比分辨率的静态图像要低。由于相机重复的安装，相机的位置有微小变化。司机是独立于本研究作者的其他人并被指示在洛杉矶大都会区社区正常行驶，选择这些区域是因为其有相对较密集的来自LAX，Santa Monica, Hollywood,Pasadena,和一些Tokyo的行人。为了移除车辆摆动的影响以此来简化标注，我们使用了Baker和Matthews的图像对齐逆组合算法来是视频稳定。

视频稳定之后，25000帧（从10小时视频中提取了137个大约几分钟长的片段）被标注了总共35000个边界框大约2300个行人。为了保证如此大规模的标注是可行的，我们构建了一个用于友好型的标注工具，如图3。该工具最显著的方面是交互过程，在该过程中标注器标注一组稀疏的帧，并且系统自动预测中间帧中行人的位置。特别的，在标注器标注了一个在最近两帧中同一个行人的边界框后，中间帧的边界框就会用三次插值发来进行插补（独立应用于边界框的每个坐标）。此后，每次标注器改变了一个边界框时，所有未被标注的帧中的边界框都会被重新插补。我们试验过更复杂的插补方案，包括依赖追踪；但是，三次插补法是最好的。对大约2.3小时的视频标注，包括验证，总过花了大约400个小时（分布在多个标注器上）。

在所有给定行人可见的帧中，标注器标记一个边界框来指示整个行人的全部范围；对于被遮挡的行人这还包括估计隐藏部分的位置。此外第二种BB用来描绘可见区域，见图五a。在一个遮挡事件中，评估的整体BB相对较为稳定而可见BB容易快速变化。比较来说，在PASCAL标注方案中只有可见BB被标注了并且被遮挡物体被标记为“截断的”。

每个单物体的一系列BB都被分配为三个标签之一。个体的行人被标注为‘Person’（大约1900个实例）。

1. **评估方法**
2. **检测算法**
3. **性能评估**
4. **讨论**