 



高斯时间感知网络(GATN)用于动作定位

# 摘要

*在视频理解中，暂时(temporally)定位视频中的动作是一个基本的挑战。大多数现有的方法往往从图像对象检测中得到启发，并扩展了SSD和更快的R-CNN等先进技术，以在ID序列中动作的暂时定位(temporal locations of action)。然而，由于预先确定的时间尺度(temporal scalse)的设计,结果可能存在鲁棒性(robustness)，这忽略了动作的时间结构，限制了检测复杂变化动作的实用性。在本文中，我们通过提出引入高斯核(Gasssian Kenel)来动态优化每个动作建议的时间尺度来解决这个问题。具体来说，我们提出了高斯时间感知网络(GTAN)-一种新的架构，它新颖的将时间结构的利用集成到一个单阶段的动作定位框架中。从技术上讲，GATN通过学习一组高斯核建模时间结构我们在未剪辑的长视频中解决时间动作定位问题。这一点很重要，因为真实应用程序中的视频通常不受约束，包含多个动作实例以及背景场景或其他活动的视频内容。为了解决这个具有挑战性的问题，我们通过三个基于片段的3D ConvNet，利用深度网络在时间动作定位中的有效性：（1）提议网络识别可能包含动作的长视频中的候选片段；（2） 分类网络学习一对多动作分类模型，作为定位网络的初始化；以及（3）定位网络微调学习的分类网络以定位每个动作实例。我们提出了一个新的损失函数的定位网络显式考虑时间重叠，并实现高的时间定位精度。最后，在预测过程中只使用建议网络和定位网络。在两个大型基准测试中，我们的方法与其他最先进的系统相比取得了显著的优越性能：在Mexaction 2上，mAP从1.7%增加到7.4%，在THUMOS 2014上从15.0%增加到19.0%。*

# 1. 介绍

最近的文献报道了动作识别方面令人印象深刻的进展[[42](#_bookmark60), [28](#_bookmark47), [2](#_bookmark20), [3](#_bookmark22), [39](#_bookmark58), [40](#_bookmark59), [24](#_bookmark43), [18](#_bookmark37), [31](#_bookmark50),[44](#_bookmark63), [13](#_bookmark32), [37](#_bookmark56)]除了在手动剪辑的短视频中检测动作外，研究人员还开始开发在野外未剪辑的长视频中检测动作的技术。这一趋势引发了另一个具有挑战性的话题——时间动作定位：给定一段未经剪辑的长视频，“特定动作何时开始和结束？”这个问题很重要，因为真实的应用程序通常涉及长时间未剪辑的视频，这些视频在空间和时间上都非常不受限制，一个视频可以包含多个动作实例加上背景场景或其他活动。在长视频（如监控视频）中对动作进行本地化可以节省大量时间和计算成本。

大多数最先进的方法依赖于手动选择的特征，它们的性能仍需要很大的改进。例如，2014年图莫斯挑战赛中表现最佳的方法[[27](#_bookmark46), [41](#_bookmark61), [17](#_bookmark36), [15](#_bookmark34)]2015年[[46](#_bookmark65), [9](#_bookmark28)]两者都使用改进的密集轨道（iDT）和费希尔矢量（FV）[[40](#_bookmark59), [25](#_bookmark44)].最近有人尝试将iDT特征与帧级深度网络自动提取的外观特征结合起来[[27](#_bookmark46), [41](#_bookmark61), [17](#_bookmark36)].然而，这种2D ConvNet不捕捉运动信息，这对于建模动作和确定其时间边界很重要。

作为静止图像中的一个类比，最近通过使用深度网络，目标检测取得了巨大的改进。受基于区域的卷积神经网络（RCNN）启发[[7](#_bookmark26)]以及它的升级版本[[6](#_bookmark25), [30](#_bookmark49), [21](#_bookmark39)]，我们开发CNN细分市场[1](#_bookmark1)，这是一个有效的时间动作定位深度网络框架，如图所示

1. 1.我们采用3DConvNets[[13](#_bookmark32), [37](#_bookmark56)]，它最近被证明有希望捕获视频中的运动特征，并添加了一个新的多阶段框架。首先，生成多尺度段作为三个深度网络的候选。提案网络将每个部分划分为行动或背景，以消除估计不太可能包含感兴趣行动的背景部分。分类网络为所有行动类别和背景训练典型的一对所有分类模型。

然而，分类网络的目的是寻找关键证据来区分不同的类别，而不是及时定位精确的动作存在。有时，分类网络的分数可能很高，即使该段与地面真相实例只有很小的重叠。这可能是有害的，因为后续的后处理步骤，如非最大抑制（NMS），可能会删除小分数段，但与地面真相有很大重叠。为了明确地考虑时间重叠，我们引入了基于相同架构的定位网络，但这一网络使用了一种新的损失函数，它奖励与基本事实有更高时间重叠的片段，从而产生更适合后期处理的置信度分数。请注意，分类网络不能被本地化网络取代。我们将在后面展示，使用经过训练的分类网络（不考虑时间重叠）来初始化定位网络（考虑时间重叠）是重要的，并且可以获得更好的时间定位精度。

总而言之，我们的主要贡献有三个方面：

（1） 据我们所知，我们的工作是第一次利用3D ConvNets和多阶段过程在野外未经剪辑的长视频中进行时间动作定位。

（2） 我们引入一个有效的多阶段分段CNN框架，提出候选分段，识别动作，并定位时间边界。建议网络通过消除不太可能出现的候选段来提高效率，而定位网络是提高时间定位精度的关键。

（3） 与适用于时间动作定位的两个大规模基准相比，所提出的技术显著优于最先进的系统。当评估中使用的重叠阈值设置为0.5时，我们的方法将Mexaction 2上的mAP从1.7%提高到7.4%，将THUMOS 2014上的mAP从15.0%提高到19.0%。我们没有对2015年图莫斯挑战赛进行评估[[9](#_bookmark28)]因为地面组织者隐瞒真相，以备将来评估。更详细的评估结果见第节[4](#_bookmark9).



1源代码和经过培训的模型可在以下网站获得：[https://](https://github.com/zhengshou/scnn/)