



基于多阶段CNN的未剪辑视频时间动作定位

郑寿、王东刚和施福昌哥伦比亚大学

美国纽约州纽约市

哥伦比亚大学{zs2262，dw2648，sc250}。埃杜

摘要

*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/)

[github。com/zhengshou/scnn/](https://github.com/zhengshou/scnn/).

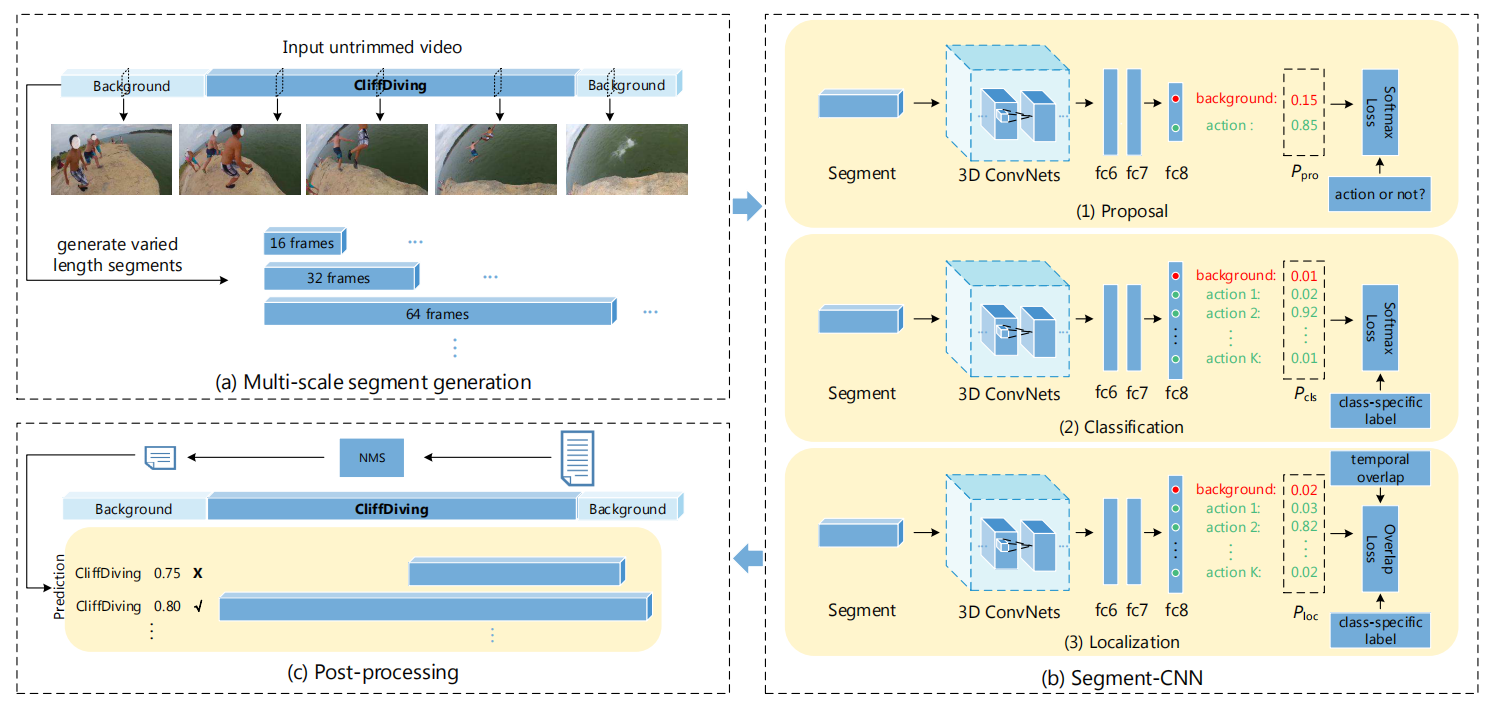


图1。概述我们的框架。

（a） 多尺度片段生成：给定未剪辑的视频，我们通过滑动窗口生成不同长度的片段；

（b） 片段CNN：proposal network识别候选片段，classification network训练动作识别模型作为定位网络的初始化，localization network及时定位动作实例并输出置信度得分；

（c） 后处理：使用来自定位网络的预测分数，我们通过NMS进一步消除冗余以获得最终结果。在训练过程中，首先学习分类网络，然后将其用作定位网络的初始化。在预测过程中，仅使用 proposal and localization networks

**2.相关工作**

**时间动作定位**。这个话题已经从两个方向进行了研究。当训练数据只有视频级别的类别标签，而没有时间标注时，研究人员将其描述为弱监督问题或多实例学习问题，以学习未经剪辑的视频中的关键证据，并通过选择关键实例对动作进行时间定位[[22](#_bookmark41), [23](#_bookmark42)]Sun等人[[36](#_bookmark55)]从网络图像中转移知识，以解决未经剪辑的网络视频中的时间定位问题。

另一个工作重点是在未经剪辑的视频（如THUMOS）中为动作实例注释时间边界时，从数据中学习。这些工作中的大多数都将其视为一个分类问题，并采用时间滑动窗口方法，其中每个窗口都被视为一个需要分类的候选动作[[25](#_bookmark44)].有关行动分类方法的调查可在[[42](#_bookmark60), [28](#_bookmark47), [2](#_bookmark20), [3](#_bookmark22)]最近，有两个方向引领了最新技术的发展：（1）Wang等人[[39](#_bookmark58)]建议沿密集轨迹提取HOG、HOF、MBH特征，然后考虑相机运动[[40](#_bookmark59)]。进一步的改进可以通过叠加具有多个时间跳过的特征来实现[[24](#_bookmark43)]（2）受最近work中CNN的成功启发[20,3]

，Karpathy等人[[18](#_bookmark37)]在大规模视频分类任务中评估帧级CNN。 Simonyan and Zisserman[[31](#_bookmark50)]设计了两个stream CNN，分别从静止图像和运动流中学习。在[[44](#_bookmark63)]提出了卷积特征映射的潜在概念描述符，并用VLAD编码实现了事件检测。为了共同学习时空特征，本文探讨了3D ConvNet的体系结构[[13](#_bookmark32), [37](#_bookmark56)]，取得有竞争力的成绩。Oneata等人[[26](#_bookmark45)]提出了近似归一化的Fisher向量来降低FV的高维性。斯托安等人[[35](#_bookmark54)]引入了两级级联以允许快速搜索操作实例。这些方法的重点不是提高精度，而是提高传统方法的效率。为了具体解决动作检测的时间精度问题，Gaidon等人[[4](#_bookmark23), [5](#_bookmark24)]用原子动作单元（Actom）对动作序列的结构进行建模。动作单元的显式建模允许匹配更完整的动作单元序列，而不仅仅是部分内容。然而，这需要对Actom进行手动注释，这可能是主观的和繁重的。本文旨在解决同样的精确时间定位问题，但不需要对原子动作单元进行手动标注。

**时空定位**

关于同时在空间和时间上定位动作，人们进行了积极的探索。Jain等人[[10](#_bookmark29)]和Soomro等人[[33](#_bookmark52)]他们的工作建立在超体素(supervoxel)上。最近，研究人员将其视为跟踪问题[[43](#_bookmark62), [8](#_bookmark27)]利用物体探测器[[11](#_bookmark30)]尤其是人体探测器[[16](#_bookmark35), [19](#_bookmark38), [8](#_bookmark27), [45](#_bookmark64)]在每个帧中检测感兴趣的区域，然后输出序列。密集的轨迹也被用来提取动作管[29,38]。Jain等人[12]添加了对象编码来帮助动作本地化。

然而，这个问题与时态局部问题不同，这是本文的主要主题。

1 当使用对象检测器寻找感兴趣的时空区域时，这种方法假设动作是由人类或其他预定义对象执行的。

2 时空定位需要在每一帧上对感兴趣的对象进行详尽的注释，作为训练数据。这使得它非常耗时，尤其是对于长时间未剪辑的视频而言，与简单标记视频中描述的动作的开始时间和结束时间的任务相比，这足以满足许多应用

**目标检测**

受深度学习方法在目标检测中的成功启发，我们还回顾了R-CNN及其变体。R-CNN包括选择性搜索、CNN特征提取、SVM分类和包围盒回归[7]。Fast R-CNN利用多任务丢失将R-CNN重塑为单阶段，并且还具有RoI池化层，用于在ConvNets中共享一幅图像的计算[6]

我们的工作与R-CNN的不同之处在于：

（1） 训练视频中的时间注释可以是多种多样的：有些是从长视频中剪下的干净修剪的动作实例，如UCF101[34]，有些是未修剪的长视频，但为动作实例注释了时间边界，如THUMOS[15,9]。我们提供了一个范例，可以处理如此多样的注释。

（2） 正如提出区域建议网络的快速R-CNN和检测对象以重新排列R-CNN结果的DeepBox所证明的，使用深度网络学习对象是有效的。因此，我们直接使用深度网络对背景和动作进行分类，以获得候选片段。

（3） 我们取消了回归阶段，因为学习视频片段的时移和持续时间回归在我们的实验中效果不佳，可能是因为动作可能非常多样，因此不包含预测开始/结束时间的一致模式。为了实现精确定位，我们使用新的损失函数来设计定位网络，以明确地考虑时间重叠。这可以降低与地面真相重叠较小的片段的分数，并增加重叠较大的片段。这也有利于后处理步骤，例如NMS，以保持更高的时间定位精度

**3.Segment-CNN的详细描述**

**3.1问题设定(Problem setup)**

**问题定义(Problem definition)**

我们将视频(video)表示为

是X中的第t帧，T是X中的帧总和。每个视频X都与一组时间动作注释相关联

M是X中action实例的总数，并且 ，分别是实例m的动作类别及其开始时间和结束时间（以帧ID衡量），，K是类别的数量。

在训练期间，我们有一个 剪辑的视频集(set)和一个未剪辑视频 集 . 每一个剪辑的视频 有 , 并且

多规模segment生成

**(Multi-scale segment generation)**

首先，每个帧(frame)都被resize成171x128像素，对于一个集的未剪辑视频X，我们进行不同长度的时间滑动窗口，如16、32、64、128、256、512帧，重叠率为75%。对于每个窗口，我们通过均匀采样16帧来构造片段。因此，对于每个未经剪辑的视频X，我们生成一组候选视频作为 proposal network的输入，H是X的滑动窗口总数， 和分别为第h个segment的开始时间和结束时间

对于一个集中的已剪辑视频X，我们直接从X匀速采样16帧的片段s

**网络结构(Network architecture)**

3D ConvNets执行3D卷积/池化，同时在空间和时间维度上运行，因此可以捕获动作的外观和运动。鉴于视频分类任务的竞争性表现，我们的深度网络在所有阶段都使用3D ConvNet作为基本架构，并遵循[37]的网络架构。所有3D池化层都使用最大池，内核大小在空间上为2×2，步幅为2，但在时间上有所不同。所有3D卷积滤波器在所有三维中的内核大小均为3，步长均为1。对3D卷积层使用符号***conv***（过滤器数量），对3D池层使用符号***pool***（时态内核大小、时态步幅），对完全连接层使用符号***fc***（过滤器数量）。在我们的架构中，这三种类型的层的布局如下所示

***conv1a*** (64) - ***pool1*** (1,1) - ***conv2a***(128) -

***pool2*** (2,2) - ***conv3a*** (256) - ***conv3b*** (256) - ***pool3*** (2,2) -

***conv4a*** (512) - ***conv4b*** (512) - ***pool4*** (2,2) - ***conv5a*** (512)

- ***conv5b*** (512) - ***pool5*** (2,2) - ***fc6*** (4096) - ***fc7*** (4096) -

***fc8*** (K + 1)

这个深度网络的每个输入是一个171×128×16的段。C3D正在Sports-1M train split[37]上训练该网络，我们使用C3D作为提案和分类网络的初始化

**3.2Training procedure**

**The proposal network**

我们训练CNN网络作为背景segment过滤器，基本上，***fc8*** 有两个节点，分别对应地表示背景（很少包含感兴趣的动作）和正在进行的动作（有很大一部分属于感兴趣的动作）

我们使用以下策略来构建训练数据

，标签

对于集中视频X的每个segment，我们把他的标签设为positive(1), 对于带有时间注释(temporal annotation)的集中视频X的候选segment, 我们通过计算其在并集（IoU）上的交集（IoU）与中的每个基本真值实例，为每个片段分配一个标签

（1） 如果最高IoU大于0.7，我们分配一个正标签；

（2） 如果最高IoU小于0.3，我们将其设置为背景

从基本事实的角度来看，如果没有与IoU大于0.7的基本事实实例重叠的段，那么如果s具有该基本事实的最大IoU且其IoU大于0.5，则我们分配一个正标签段s

最后，我们得到了，它由所有 positive片段和随机抽样背景段组成，其中

在所有实验中，除了***fc8*** 为0.01、动量为0.9、权重衰减因子为0.0005外，我们使用的学习率为0.0001，并且每10K迭代将学习率降低10倍。总迭代次数取决于数据集的规模，将在第4节中阐明。

请注意，与多类别分类网络相比，该提议网络更简单，因为输出层仅由两个节点（动作或背景）组成

**The classification network**

在proposal network移除大量背景片段后，我们为K个action类别和背景训练一个分类模型。

准备训练数据遵循proposal network的类似策略

除为positive指定标签外，分类网络明确指示(action category)行动类别。此外，为了平衡每个类的训练数据数量，我们将背景实例的数量减少到

对于SGD中的参数，学习率为0.0001，fc8为0.01，动量为0.9，权重衰减系数为0.0005，每10K迭代一次，学习率除以2，因为随着类数的增加，收敛速度会变慢

**The localization network**

如图2所示，重要的是提高与地面真相实例重叠较大的段的预测分数，并降低重叠较小的段的分数，以确保后续后处理步骤可以选择重叠较大的段，而不是重叠较小的段。因此，我们提出了一种具有新损失函数的定位网络，它考虑了IoU和地面真实情况。

定位网络的训练数据通过将每个片段与重叠测量相关联，从Scls中增加。具体而言，我们将修剪视频中的设置为=1。如果来自未剪辑的视频，并且具有正标签k，我们将设置为等于片段与相关地面真相实例的重叠（通过IoU测量）。如果是一个背景段，我们可以在后面看到，它的重叠测量不会影响我们在反向传播中的新损失函数和梯度计算，因此我们只需将其设置为1

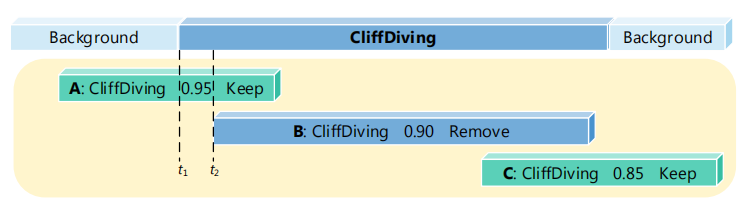


Figure2。典型的错误定位案例。假设系统输出三个预测：A、B、C。可能是因为[t1、t2]期间存在一些证据，A的预测得分最高。因此，NMS将保留A，删除B，然后保留C。然而，实际上我们希望在NMS中删除A和C，并保留B，因为B具有最大的与地面真相实例相关的IoU

在每个小批量中我们有N个训练样本

对于第n个segment，***fc8*** 的输出向量为，softmax层之后的预测得分向量为。注意对第i个类别(class),

新的损失函数由 和 组成

(1)

平衡了各部分的贡献，通过实证验证，我们发现=1在实践中效果良好。Lsoftmax是传统的softmax损耗，定义为

(2)

这对于训练深层网络进行分类是有效的

旨在联合减少分类错误，并根据重叠程度调整置信度得分的强度

(3)

这里，当真实类标签为正时，[>0]等于1，当=0时等于0，这意味着是背景训练样本。旨在提高与地面真相实例高重叠部分（）的检测分数（），并降低小重叠部分的分数。超参数控制置信度得分强度的调整范围。第4节探讨了的敏感性。此外，第个节点的总梯度w.r.t输出***fc8***如下：

(4)

in which

and

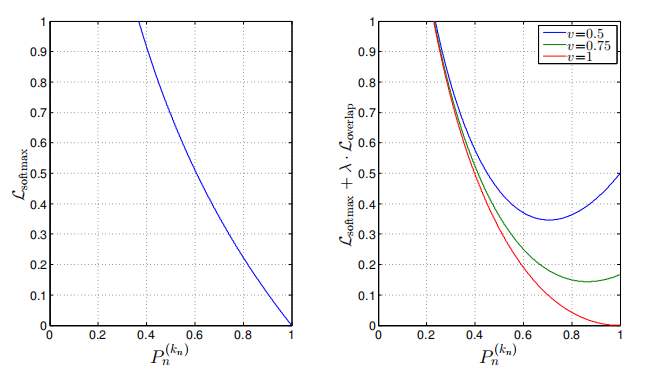


图3。说明与在每一个正向片段中的工作原理。这里我们使用=1，=1，并在中改变重叠。轴是与真标签对应的节点处的预测分数，轴是损失

给定一个训练样本,图3显示了如何影响最初的softmax损耗,它还提供了关于这种损失函数设计的更具体的见解。

1. 如果该段属于背景，则，。
2. 如果该段为正值(positive),则在处达到最小值,因此会惩罚两种情况

要么由于错误分类而太小，要么爆炸并超过学习目标P，该学习目标与重叠成比例

还要注意，的设计是随着的减少而增加

考虑到与地面真实情况重叠较小的训练段可能包含不可考虑的噪声，因此可靠性较低

1. 特别是，如果该正段(positive segment)具有重叠=1，则损耗函数变得类似于softmax损耗，且 从+∞ 到1随着从0变为1时。在优化过程中，会在上进行微调。因为分类也是定位网络的一个目标，而经过训练的分类网络可以很好地初始化。我们使用与分类网络相同的学习率、动量和权重衰减因子。取决于数据集的其他参数见第4节

**3.3Prediction and post-processing**

预测和后处理

在预测过程中，我们滑动不同长度的时间窗口来生成一组片段，并将它们输入以获得建议分数。在本文中，我们使用≥ 0.7. 然后，我们通过评估保留片段，以获得动作类别预测和置信度得分。在后处理过程中，我们删除所有预测为背景的片段，并通过乘以训练数据中每个窗口长度的特定于类的出现频率来优化，以利用窗口长度分布模式。最后，由于评估中不允许冗余检测，我们基于进行NMS以去除冗余检测，并将NMS中的重叠阈值设置为略小于评估中的重叠阈值θ（θ− 0.1）。

的是提高与地面真相实例重叠较大的段的预测分数，并降低重叠较小的段的分数，以确保后续后处理步骤可

**4.Experiment实验**

**4.1Datasets and setup**

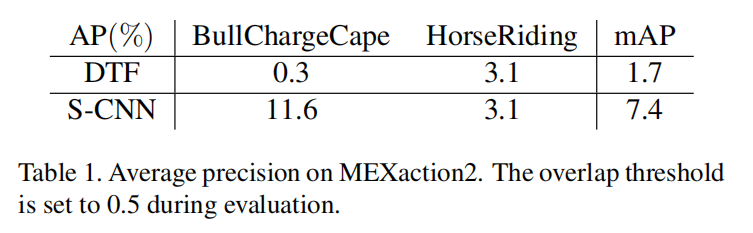
MEXaction2[1]。这个数据集包含两个动作类：“斗牛”和“骑马”。该数据集由三个子集组成：INA视频、YouTube视频和UCF101骑马视频。YouTube视频和UCF101骑马视频被剪辑，而INA视频则没有剪辑，总共大约77小时。对于带有时间注释的动作实例，它们被分为训练集（1336个实例）、验证集（310个实例）和测试集（329个实例）

THUMOS 2014[15]。THUMOS Challenge 2014中的临时动作检测任务致力于在长时间未剪辑的视频中定位动作实例。检测任务涉及20个类别，如图4所示。用于培训的剪辑视频是UCF101中这20个动作的2755个视频。验证集包含1010个未经修改的视频，共有3007个实例的时间注释。测试集包含1574个未剪辑视频中的3358个动作实例，而其中只有213个包含感兴趣的动作实例。我们排除了测试集中剩余的1361个背景视频。

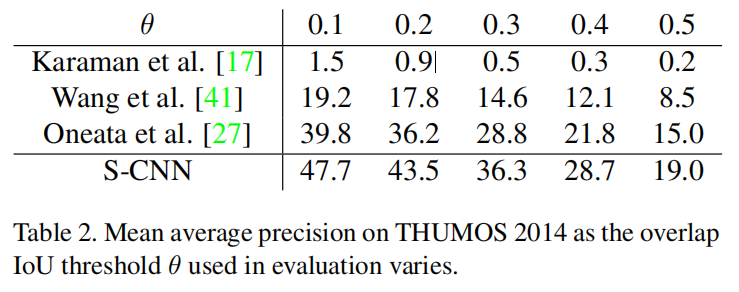
**4.2. Comparison with state-of-the-art systems（与最先进系统的比较）**

**评估指标**。我们遵循THUMOS Challenge中使用的常规指标，将时间动作定位视为检索问题，并评估平均精度（AP）。只有当预测具有正确的类别预测，且具有大于重叠阈值（通过IoU测量）的地面真相实例的IoU时，预测才会被标记为正确。请注意，不允许进行冗余检测。

**MEXaction2的结果**。我们基于Caffe[14]和C3D[37]构建了我们的系统。我们使用MEXaction2中的训练集进行训练。建议网络的训练迭代次数为30K，分类网络的训练迭代次数为20K，本地化网络的训练迭代次数为20K，α=0.25。我们使用上述设置S-CNN来表示我们的CNN片段，并与典型的密集轨迹特征（DTF）和视觉文字袋表示进行比较。DTF的结果由[1]提供，它使用不同的负样本集训练三个SVM模型，并总体平均AP。根据表1，我们的细分市场CNN在“斗牛”动作和“骑马”动作中获得了巨大的性能增益。图5分别显示了我们对“斗牛角”和“骑马”的预测结果。



**THUMOS 2014上的结果**，训练集和验证集中的实例用于训练。所有三个网络的训练迭代次数均为30K。我们再次为本地化网络设置α=0.25。我们使用上述设置将我们的CNN片段表示为S-CNN



至于比较，除了DTF，还有几个基线系统包含帧级深度网络，甚至还利用了许多其他功能：

（1） Karaman等人[17]使用基于加权显著性池的iDT FV编码，并使用帧级CNN特征进行后期融合。

1. Wang等人[41]在iDT上构建了一个具有FV表示和帧级CNN特征的系统，并进行后处理以细化检测结果。

（3）Oneata等人[27]在时间滑动窗口上使用iDT的FV编码进行定位，并执行后续处理[25]。最后，他们对时间窗口的定位分数和根据iDT特征、图像特征和音频特征训练的分类器生成的视频级别分数进行加权融合。结果如表2所示。图4显示了每个类的AP。我们的细分CNN在20个动作中有14个明显优于其他系统，平均性能从15.0%提高到19.0%。我们还在图6中展示了THUMOS 2014测试集的两个预测结果

**效率分析**。与所有其他系统相比，我们的方法非常有效，这些系统通常会融合不同的功能，因此可能会变得非常麻烦。滑动窗口生成的大部分片段被第一提议网络删除，从而大大减少了分类和定位操作。对于每个批次，速度约为1秒，每个批次中可处理的段数取决于GPU内存（对于4G内存的GeForce GTX 980，大约25段）。存储需求也非常小，因为我们的方法不需要缓存中高维特征，例如FV来训练SVM。CNN所需的全部是三种深度网络模型，它们总共占用不到1GB

**4.3. Impact of individual networks**

为了单独研究每个网络的影响，我们使用不同的设置比较了四段CNN：

1. S-CNN：保留第4.2节中的所有三个网络和设置，并且在上进行微调；
2. S-CNN（无提案）：完全移除提案网络，直接使用在滑动窗口上进行预测；
3. S-CNN（无分类）：完全移除分类网络，因此没有作为培训初始化的分类网络；
4. S-CNN（无本地化）：完全移除本地化网络，而是使用分类模型生成预测。

**提案网络**。我们比较了S-CNN（无提案）和S-CNN，后者包括上述提案网络（fc8中的两个节点）。由于S-CNN比S-CNN（w/o提案）的网络结构更小，S-CNN可以减少在背景段上进行的操作数量，从而加快速度。此外，表3中列出的结果表明，保留提案网络也可以提高精度，因为它是为过滤掉缺乏兴趣行动的背景片段而设计的。

**分类网络。**虽然在预测过程中没有使用，但分类网络仍然很重要，因为对进行微调可以获得更好的性能。在评估过程中，我们对最终预测结果进行top-κ选择，以选择置信度最高的κ片段。如图7所示，当κ变化时，在上微调的S-CNN始终优于S-CNN（w/o分类），因此分类网络在训练期间是必要的

**本地化**localization**网络**。图7也证明了本地化网络的有效性。通过添加定位网络，S-CNN可以比基线S-CNN（w/o本地化）显著提高性能，后者只包含建议和分类网络。这是因为在定位网络中引入的新损失函数细化了与地面真相有较高重叠的偏好段的分数，因此可以实现更高的时间定位精度

此外，我们在损失函数的重叠损失项Loverlap中改变α来评估其灵敏度。我们发现，我们的方法在α值范围内（例如，从0.25到1.0）具有稳定的性能

**5.结论**

我们提出了一个称为分段CNN的有效多阶段框架来解决未剪辑长视频中的时间动作定位问题。通过对每个网络的上述评估，我们展示了提案网络对识别候选细分市场的贡献，以及分类网络为培训本地化模型提供良好初始化的必要性，以及在定位网络中使用的新损失函数在时间上精确定位动作实例的有效性。在未来，我们希望将我们的工作扩展到通常由多个动作组成的事件和活动，因此及时准确地定位动作实例有助于识别和检测它们。

**6.致谢**

这项工作由情报高级研究项目活动（IARPA）通过内政部国家商业中心合同编号D11PC20071提供支持。美国政府有权出于政府目的复制和分发再版，尽管有任何版权注释。免责声明：本文中包含的观点和结论是作者的观点和结论，不应被解释为必然代表IARPA、DOI-NBC或美国政府的官方政策或明示或暗示的支持。我们感谢刘东、叶光南和匿名评论者提出的富有洞察力的建议