指导教师： 杨涛　　提交时间： 2016.3.13

CVPR2015 Paper Translation

No: 59

姓名： 成舜

学号： 2013302570

班号： 10011305

人类行为与分层分割超体元一致性

Jiasen Lu1, Ran Xu1 and Jason J. Corso2
1 计算机科学与工程,纽约州立大学布法罗分校
2 电气工程和计算机科学,密歇根大学

概要

详细分析人类行为,如动作分类、检测和定位从社区得到了越来越多的关注;数据集JHMDB等合理的进行研究分析了影响这样的更深层次的信息在更大的行动的理解问题。然而,详细的自动分割的人类行为相对未开发。在本文中,我们在这个方向迈出一步,提出一种分层 MRF 模型桥低级视频片段与高级人类运动和外观;新的高阶势连接不同级别的超体元层次执行的一致性人类分割不同尺度段。我们的单层模型明显优于当前最先进的动作模型并且我们的完整模型改进了单层基线行为细分。

1.介绍

近年来,一个伟大视频理解的突破已经被运用于行动识别(33,28，17)等大型数据集 UCF101[30]和 HMDB51[20]。对视频进行分类,提出了表征,从低级的特性,利用点运动轨迹和本地外观/信息(33、34),高级功能,创建一个高维的行动空间[28],利用人类[32],以及深层神经网络的无监督特征[17]。

尽管进展慢,然而,这些方法仍然有限的能力提供任何比视频教程行动更深的信息标签。事实上,已经开始关注视频理解等问题例如分类群活动[22]和[11]人机交互。然而,这些焦点仍很粗糙并且不适合许多应用,如自主驾驶[23]机器人手术[4],需要精确的行动在时空界限。

相比之下,行动本地化(6、21)和行动直接检测[31]更加强调特定兴趣行动发生的"时间"和"地点"在视频中。这些美好的行动不仅推断使一组广泛的应用程序,Jhuang 等。[16]最近还显示,精确的人体轮廓边界分类本身会影响行动。他们最近评估的影响,不同的真实场景操作分类;他们详细分析注释的人类行为的数据,发现人类"傀儡"提供了重要的帮助对更好的行动分类与整个视频甚至边界框约束人类行为。然而,他们不构成解决自动本地化和自动分割轮廓或动作的方案。

一个方向的自动行动改进分割遵循范式模板,通常是在一系列的边界框的形式或边界体积的形式,用于本地化行为在时空中。模板是刚性的,手动选择[8 6]或变形与灵活的属性(31,35)。另一个方向是由低级分割,如手工剪裁人类段[18],管让[14],合并超体元[9，36]使用运动和时空段[24]寻求类似人类的部分用颜色,形状和运动线索。最新的结果表明,两个方向(35,14)实现先进的性能,同时[35]使用拥有更强的监督,如给训练带来注释,这是不平凡的。

利用低级分割如超体元[9，36]行动分割的前兆,定位和分类是一种很有前途的方向。它可能放松所需的监督并提供普遍表示框架。然而,有一些需要解决的挑战。首先,根据研究 Jhuang 的研究。[16],整个人类分割比占一个边界框组成的块段 1 如图 1。人类行为细分的系统输出,J-HMDB 数据集的例子是使用原始 RGB 呈现分割区域像素为前景和背景区域覆盖着一个透明的黑色模型。

例如[24]。第二,分割方法只取决于运动,例如[14],往往错过对象或人类的部分是静态的视频。例如,目前尚不清楚人们小时候向桌子挥舞的手是全部或部分分割前景。第三,进一步细分质量是至关重要的视频理解途径,从而使分割粒度的选择变成一个简单的问题(37 岁,26)所示。

为了这些目的,我们提出了一种分层 MRF 模型人类行为细分,满足以下目标。

•自动段整个人类行为轮廓,如图 1 所示从而进一步使更深层次的视频理解任务,即动作分类和边缘本地化

（接左下）

•连接低级分割和高级人类恢复静态身体部位和困难,阐明身体部位。

•提高这些方法的信息质量执行超体元不同尺度之间的一致性(水平)的层次结构。

1.1 我们的方法概述

我们首先提出一个人类运动特点，它表示能够占到摄像机运动和平衡人体运动和人类自动出现提示。类似的概念 chen 提出的称为"动作模型"等。[5]:它产生一个排序的视频区域根据它们包含一个行动的程度,但该地区排名是小 3 d cuboid-volumes(见图 2)和真实是诺斯替人类行为的界限。人类的运动特点,相反,人类行为轮廓自然结合。具体而言,我们估计前景运动形成一个相机模型通过长期轨迹[3],并获得一个人类之前由 DPM-based 得到的人模型[7]。我们比较人类的运动特点和基于光流的摄像机运动估计方法[25]和动作模型[5],并找到一个+ 16%相对改善动作模型排名。(见 4.1 节.更多细节。)

部分人类行动,开始运用分层的基于视频分割[38]超体元层次。在这个层次,我们定义一个 MRF 模型,使用我们的新人类运动特点如 Jain 等[15]指出,与广大 temporal-extent 方差能使图的超体元脆弱。而不是选择基于帧的超体元图,可能会失去人类行为边界,我们另外设计一个成对潜在基于邻近超体元连接只在光流的方向。因此,我们暂时只考虑成对的超体元。另一方面,为了解决这一问题的静态人体部分,我们从学习中提取一个形状之前的部分 person-DPM[7]和信用卡诈骗罪 rconnections 超体元和形状之间之前的一对巨大的潜力。

在这个分层 MRF,我们设计一个创新的高层次的不同超体元之间潜在的不同级别的层次结构。凭直觉,超体元高水平的层次进行更好的人类或人类的部分语义,但更容易受到相应的泄漏。同样,超体元在低水平减少泄漏,还携带语义。大多数现有方法手动选择层次水平基于视觉检查和最佳选择层次水平是基于不同的视频,Oneata 等。[26]确定层次水平的超体元通过寻找最好的本地化 xu 提出的分数训练集等。[37]提出一个从不同层次压扁过程细分选择超体元可以使用统一的熵判据。缓解泄漏和维持更好的语义信息,我们的高阶超体元潜在好处是可以从更高层次约束运动和外观线索。解决人类的运动分割问题,我们发现它有效地定量和定性实验获得.见 4.2。

最后,我们的能量最小化的分层 MRFα-expansion 算法(1、19)图 2 和现在的一个方法。运动特点和人类卓越特性。(a)原始图像。(b)的可视化人类卓越的回应。2(c)初始集群的轨迹 GMM 前景和蓝色背景(红色)。(d)的可视化运动显著响应轨迹分类错误(注意,(c)的低反应)。(e)的可视化前景运动与光流估计,使用[25.14.13]。(f)动作模型 5

自动学习基于 GMM 估计模型参数。剩下的部分组织如下。部分 2 和 3 详细制定困难曼提出了我们的模型。第四部分提出了定量和定性的评估方法。第五节总结本文并讨论未来的工作。

2。人类运动特点对人类行为细分

我们的方法输入包含人类行为的视频和用输出的视频时空分割标签所有人类行为作为前景,或者背景像素。我们不假设场景上下文或人类的水平清晰度;我们因此使用数据集,如 JHMDB[16],UCF-sports[27],佩恩行动[39]因为它们含有人类行为与大变异,总变形和强大的相机运动。详细的注释,如人类木偶或造成关节可用于详细评估。

我们开始讨论方法和人类运动特点自动测量的新方法。这个新特性可以直接用于本地化和人类行为,我们直接评估(4 节),它可以用作一个动作之后建模功能,正如我们在 3 节。

人类运动特点包含两个部分:前台运动和人类外表的信息。对前景运动估计,建立了一个自然的选择。目前的前景评估方法一般分为两类,1)使用光流和 RANSAC 找到主导运动作为背景运动[29],2)光谱聚类与长期的轨迹,找到主要轨道集团[2]。我们早期的实验表明光学流可能是不可靠的,见图 2(e)为例,并且集群的轨迹,虽然健壮,有时也可能离开离群值,如图 2 中的红点在后台(c)。我们因此把这两个方案,提出一种新的运动卓越特性,我们使用长期轨迹建立相机运动模型,然后测量运动

通过从相机模型偏差显著。我们使用一个 2 d 摄像机运动参数仿射运动模型。具体地说,给定一组轨迹与 L Tr 的视频帧。两个轨迹之间的速度差异 trj 在时间 t 是

$$d_t(tr^i, tr^j) = \frac{1}{T}(u_t^i - u_t^j)^2 + (v_t^i - v_t^j)^2 \qquad (1)$$

在 ui t 和 vj t 表示三聚合的帧的运动。我们测量 trj 使用速度的中值三之间的距离和其他所有的沙子进一步符合一维高斯混合模型得到两个集群如[10].

然而,没有结构信息,后台轨迹都被分组在前台集群。减轻假聚类,我们计算仿射运动模型和适应轨迹点与背景集群上的规则

$$\hat{\theta} = \arg\min_{\theta} \sum_{p \in \Omega} \rho(r_{\theta}(p,t)) \qquad (2)$$

θ 是仿射运动模型参数,ρ(·)的定义是健壮的图基函数[12]和 rθ(p、t)是流离失所的轨迹点 p 帧 t 框架区别。我们进一步调整所有的轨迹的运动特点计算平均偏差通过剪辑。我们可以看到在图 2(d),(c)的分类错误的轨迹是评审的背景。人类外观信息,我们使用一个 DPM[7]人检测器训练在 2007 年帕斯卡 VOC 和构造一个显著图标准化检测分数的平均规模和所有的组件,如图 2 所示(b)。我们进一步编码前景运动和人类在超体元凸起了 3.1 节。

3. 分层 MRF 图结构

在本节中,我们介绍我们的分层马尔可夫随机场(MRF)超体元模型的层次结构。我们表示图 G nodesX 和边组成的 E,其中 X 是超体元集在整个视频体积和 E 的边缘设置,如图 3 所示。在证券交易委员会讨论。1,与所有超体元建立一个图表

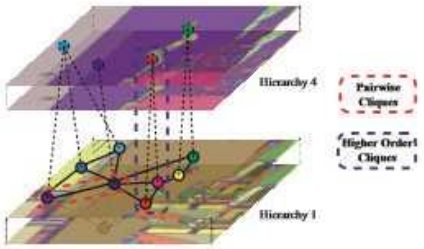


图 3。提出了分层 MRF 图。节点超体元中如果两个候选人边缘存在超体元是邻居,但是确认只有通过 s 超体元探测器光学流或重叠的人。高阶派系之间是由相应的高级超体元水平。为描述简单起见只有一小部分节点连接。

视频可能由于大尺寸的邻居导致脆弱的图。不像[15],从每一帧到超体元建立图形和传播后续的框架,我们设计一种机制来约束图的边的数量。凭直觉,我们只建立一个边缘两个超体元(xi,xj)1)他们在光流的方向是邻居,邻居暂时发现两个超体元,或者 2)两个超体元都重叠于一个人的检测,有信心值大于某个阈值,这在很大程度上发现两个邻居超体元空间即使受制于人类的外表。

红色虚线内的节点在图 3 中表示一个证实了边缘。更详细的报告.我们使用 V 表示的集合超体元更高一层的层次结构。v∈每个元素代表一个高阶集团(蓝色虚线表示图 3 中的信件)。而且,青年志愿表示标签的集合分配给超体元节点属于超体元告诉我们一个随机变量易建立∈{ + 1−1 }与每个节点代表标签可能需要,从而可以 human-with-action(+1)或背景(−1)。我们的目标是将所有的超体元在整个视频中表示 X = { 11 } N i = 1。

3.1。能量函数

鉴于图结构 G =(X,E)诱导的超体元层次图中(E 是这些 fedges 层次)。我们引入一个能量函数/ G =(X,E),实施分层超体元通过高阶势源于超体元 V 一致性。

$$E(Y) = \sum_{i \in \mathcal{X}} \Phi_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \Phi_{i,j}(y_i, y_j) + \sum_{v \in \mathcal{V}} \Phi_v(y_v) \qquad (3)$$

Φi(yi)表示一元超体元与潜力指数 i,Φi,j(咦,yj)表示 spairwise 两个超体元之间潜在的优势,和 Φv(青年志愿)表示两层之间的高阶超体元的潜力。

一元可能性:

我们编码运动特点和人类卓越特性得到一元潜在的组件:

$$\Phi_i(y_i) = \gamma_M M_i(y_i) + \gamma_P P_i(y_i) + \gamma_S S_i(y_i) \qquad (4)$$

γM,γP 和 γS 权重为一元。Mi(yi)反映了运动的证据,π(yi)和 Si(yi)分别反映了人类的证据。Mi(yi)可以计算为:

$$M_i(y_i) = \exp\left(-\frac{\lambda_m}{|x_i|} \sum_{tr^j \in x_i} w^M(tr^j)\right) \qquad (5)$$

λm 尺度参数,wM(trj)运动轨迹 trj 凸起体重和|xi| supervoxel xi 的大小。

人类的卓越 π(yi)可以形成:

$$P_i(y_i) = \frac{1}{1 - \exp(-\frac{\lambda_p}{|x_i|} \sum_{px^j \in x_i} w^P(px^j))} \qquad (6)$$

wP(pxj)是人类体重显著像素 j。我们使用金刚石与模特培训在 2007 年帕斯卡 VOC 检测人类发现和形成规范化的显著图平均检测成绩的尺度和组件。

[24]后,背景物体的直边界是常见的人造场景,但人体边界包含更少的零曲率点。所以我们也计算曲率边界点的每个超体元凸起的形状特征。

$$S_i(y_i) = \exp\left(-\frac{\lambda_s}{|x_i|} \sum_{px^j px^k \in B_i} w^S(px^j, px^k)\right) \qquad (7)$$

Bi 是边界点的集合在超体元 xi,(pxj pxk)是两个附

近的像素和 wS 曲率。

成对潜力:如秒。3 所述,我们限制边缘空间只有两种类型的邻居:时间超体元邻居和人类 appearance-aware 空间邻居,所以我们定义成对势为:

$$\Phi_{i,j}(y_i, y_j) = \gamma_I I_{i,j}(y_i, y_j) + \gamma_K K_{i,j}(y_i, y_j)) \quad (8)$$

γI 和 γK 成对潜在的权重。I，i,j(yi,yj)之间的成本是超体元 I 和超体元 j 的人类检测约束,确保平滑

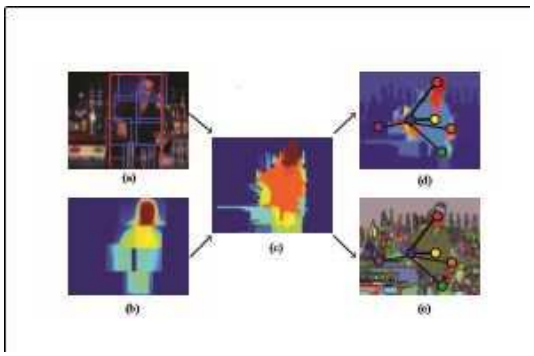


图 4. 人类用行动表示(一)DPM 检测与根和边界框的一部分。(b)相应的 DPM 面具从 PASACALVOC 中提取一部分。(c)超体元响应模型。(d)和(e)超体元成对连接的运动特点分别映射和分割,粗线表示强烈的空间联系。

注意,我可以确定和 j 作为邻国没有进行像素级连接。Ki j(yi,yj)是确保平滑暂时虚拟不同。部分原因模型的混合组件通常反映出人类的身体部位。我们利用这个通过定义新的潜力,利用前一个形状为每个部分。我们将这些信息在我们的模型中通过突出超体元在同一个人类探测与强大的共识。因此,我们定义 rt 我超体元的反应检测在 t 帧:

$$r_i^t = s_r^t \mu_r^t \max(\mu_p^t) \quad (9)$$

的大小。μt r 是百分比.超体元(我们使用 2 d 超体元平面)位于边界框和 μt p 是百分比超体元在于模型从 PASCALVOC 中提取一部分,seeFig 4(b)的一个例子。人类连接分两个超体元我和 j 在帧 t Sti,j = min(rt,rt j)。因此我们定义

$$I_{i,j}(y_i, y_j) = \delta(y_i \neq y_j) \exp(-\beta_I \sum_{t \in T} S_{i,j}) \quad (10$$

在 βI 尺度参数。图 4 描述了这个过程。颞平滑项 Ki,j(yi,yj)被定义为

$$K_{i,j}(y_i, y_j) = \delta(y_i \neq y_j) \exp(-\beta_D D(x_i, x_j)) \quad ($$

D(xi,xj)是两个超体元 χ2 的直方图特征之间的距离。对于每个超体元我们计算出两个特点:1)一个 RGB 颜色直方图与 33 箱(11 箱每通道),和 2)直方图光学流 9 箱。

高阶潜力:我们定义层次超体元标签一致性的潜力。不同于[15],使用高阶势超体元时间平滑,我们利用不同超体元层级之间的连接。在实践中,我们采用健壮的 Pn 模型[19]定义的潜力

$$\Phi_v(y_v) = \begin{cases} N(y_v)\frac{1}{Q}\gamma_{\max}(v) & \text{if } N(y_v) \leqslant Q \\ \gamma_{\max}(v) & \text{otherwise} \end{cases}$$

青年志愿表示标签的所有节点对应于高级超体元层次 v∈诉 N v(q)内的节点数量不占主导地位的标签。w 是一个截断参数,严格控制我们想要执行超体元在两层内的一致性。

法则 γmax(v)是一个函数考虑大小,颜色和运动潜水员超体元爆破。如果前台超体元和背景超体元不当合并在 v,γmax(v)应该大,少惩罚是买的标签不一致。具体来说,γmax(v)= | |电视上学到 exp(−η(σc v +σm v)),σc v 和 σm v 是 RGB 颜色的方差以及运动超体元 v。

3.2。能量最小化 Eqn。

3 中定义的能量函数和参数可以有效地使用 α-expansion 最小化算法[19]。地方色彩是分割的强有力的证据。因为我们没有最初的标签帧,我们把通过学习高斯混合模型（GMM）R G B 第一次使用的输出分割结果进一步完善成颜色空间。

γM 的价值,γPγS 反映了体重运动和人类的线索。对于大多数视频,我们设置了一元权重合理价值,γM = 6,γP = 4 和 γS = 3 在我们的实验。然而,对于视频很少运动,如高尔夫球,我们希望人类主宰一元项检测功能。因此,我们自动确定 γM 和 γP 通过比较的均值估计高斯中心和所有个人距离的意思。我们实证发现 GMM 估计性能更好。但 GMM 估计会失败如果 μ2 位于一个局外人。在这种情况下,两个估计的值明显不同。如果大于一个阈值的差异,我们设置 γM = 3,γP = 7。

我们设置了重量参数 γI =γK = 0.1,γH = 2。规模参数 λmλp、βIβD 和 η 都将自动设置为意味着所有个人距离的倒数。截断参数 Q = 0.3 |γv|。我们使用相同的参数设置在所有的数据集。

4.实验

全面评估我们的方法,我们报告结果在三个任务:首先,与我们人类运动卓越特性,我们评估动作模型排

名和比较先进的方法在 4.1 秒。(5、25);第二,我们评估人类行为与基线方法分割,没有成对或高阶项,在秒。4.2;第三,不把分割出来,我们评估的行动分类和定位与其他先进的方法,在 4.3 秒。。数据集
前面介绍的一样,我们的方法并没有假设场景上下文,相机运动或水平清晰度的人类,所以一般行动识别或本地化数据集"野生"是合适的。此外,由于我们的方法分割的本质,我们赞成与地面实况数据集分割,如。,JHMDB[16],或至少与地面边界框真理,如 UCF-Sports[27]和佩恩行动[39]。
UCF-Sports 包含 150 个视频超过 10 类有大量人类行为变异,总变形和强大的相机运动,我们充分评估动作模型排名,动作分割、分类和定位的数据集,因为它是广泛使用在社区和评估。我们使用培训/测试分裂为这些任务[5]。
JHMDB HMDB 的一个子集,包含 928 个片段组成的 21 个动作类别,所有帧都带注释的发出"人类傀儡",我们拿一个酸的人类行为分割地面实况进行评估。我们评估我们的行动分割和动作识别与这个数据集。佩恩动作包含 15 和 2326 视频剪辑

4.1。评估动作模型

[5]后,我们测量意味着动作模型排名的平均精度(mAP)运动显著图和联合人类运动的显著图,描述秒。2,并比较与[5]使用 ranking-CRF 等级 3 d 的动作模型长方体卷,和运动 2 d[25]

	subset of video				all video		
	[5]	[25]	Motion	Joint	[25]	Motion	Joint
dive	58.7	63.9	**69.5**	66.4	58.1	**66.7**	64.1
golf	61.8	41.9	66.6	**66.8**	35.2	63.7	**69.1**
kick	68.7	71.1	**73.0**	61.3	70.7	**67.9**	60.7
lift	**86.7**	61.5	76.4	77.1	67.2	**76.1**	75.4
ride	18.8	52.4	51.3	**52.5**	33.1	35.7	**42.3**
run	48.4	48.2	**61.6**	61.0	51.5	58.4	**59.1**
skate	57.6	65.4	**81.1**	63.8	55.4	57.7	**66.6**
sw-b	80.1	**89.8**	87.3	84.4	**85.7**	80.9	77.3
sw-s	54.0	64.9	78.9	**79.5**	61.8	**72.7**	67.0
walk	50.4	55.3	70.3	**79.1**	40.5	57.1	**69.1**
Avg.	60.8	61.9	**71.8**	68.6	56.0	63.9	**65.4**

Table 1. mAP of our actionness with proposed motion and joint feature against actionness [5] on UCF sport dataset

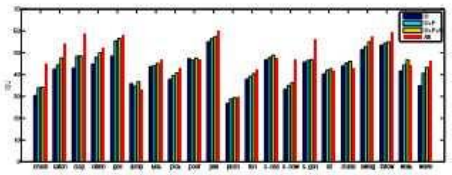


图 5

行动与 JHMDB 借据测量数据集分割结果。我们报告三个基线和模型在所有 21 类。产生前景运动特点与光流和 RANSAC 地图。从表 1,我们可以看到人类运动显著地图显示了超过 10%的显著改善地图获得先进的方法[5]。此外,我们的方法的需求非常有限的监督与帕斯卡 VOC(只有基于图像的人模型训练),这使得一个好的潜力进一步详细行动理解任务。我们还观察到整个 UCF-Sports 数据集的分数远低于测试集,这可能表明,测试集人类行为有更多的线索和低噪音而训练集。

4.2。人类行为细分

我们彻底评估人类行为细分方法这三个数据集和所有基本线,我们表示你只使用一元的潜力,U + P 使用一元,成对的潜力,U + P + H 一样使用额外的高阶势和所有我们的完整模型。表。2 总结了借据,平均精度和平均回忆,它演示了成对潜力一般贡献 1%−3%获得一元的潜力,和高阶术语我们观察另外 1%−3% / U + P .令人惊讶的是,第二个路径推理与额外的颜色 priorfrom U + P + H 显示另一个 1%−4%的涨幅。比分演示了我们的模型有效性。定性,图 6 显示了地面真理,一元分割掩模和完整的模型,完整的模型可以有效地去除 mis-segmented 一元的地区分割,例如"跳水"、"骑着马"和"运行",由于成对和高阶一致性。此外,我们完整的模型也可能完成缺失段与一元分割,如"高尔夫球"和"踢",可能是因为 human-aware 边缘的成对的潜力。InFig。来说 5,我们将展示在 JHMDB 数据集分割结果,行动完全直立人体姿势,如"高尔夫球"和"拉动式"执行相对比操作和大量运动关节,如。"跳",或人类的运动更从小孩而不是成人,例如"推"。定性,在图 6 中,我们自动生成面具使很好与地面实况面具。我们的一些面具泄漏由于超体元细分如"选择"、"推"和"坐",但值得注意的是,随着地面实况与变形带注释的"傀儡",对齐可能不是完美当人体不能适应傀儡如"波"和"推"。在这些情况下我们的方法生成覆盖整个人体的模型。

	UCF-Sport	JHMDB	Penn Action
Baseline DT	83.0	54.4	94.5
Seg DT*	**93.6**	**58.6**	95.0
GT bbox DT	89.0	55.5	**95.4**
GT puppet mask DT	-	56.2	-

Table 3. Action Recognition results

	subset of frames			all frame	
	[21]	[24]	Ours	[24]	Ours
dive	43.4	46.7	**48.0**	44.3	**48.3**
golf	37.1	**51.3**	49.6	**50.5**	50.0
kick	36.8	**50.6**	36.0	**48.3**	35.5
lift	**68.8**	55.0	57.2	51.4	**57.1**
ride	21.9	29.5	**29.8**	**30.6**	29.8
run	20.1	**34.3**	33.9	33.1	**33.7**
skate	13.0	40.0	**46.1**	38.5	**45.9**
swing-b	32.7	54.8	**62.6**	54.3	**62.3**
swing-s	16.4	19.3	**53.9**	20.6	**54.9**
walk	28.3	39.5	**58.1**	39	**58.1**
Avg.	31.8	42.1	**48.1**	41.0	**48.0**

Table 4. Action localization results measured as average IOU (in %) on the UCF sports dataset

4.3。动作识别和定位

在得到人类行为细分后,我们进一步评估如何影响行动的识别和定位。首先,我们提取密集的轨迹特性在整个视频,2)我们的分割掩模,3)地面真理边界框和 4)地面真理木偶面具 JHMDB(只)。我们发现使用边界框或分割掩模通常可以提高分类精度在整个视频,特别是使用分割掩模通常比边界框得到更好的结果。[16]的发现是一致的评估和证明我们的方法是一个似是而非的工具更好的视频理解。值得注意的是,我们 UCF-Sports 分类精度达到+ 10%的涨幅近有效视频理解论文[24](81.7%)和[14](80.24%)。

此外,在 JHMDB 数据集,我们发现使用分割掩模实现更好的准确性比地面实况木偶面具,也发现在 4.2 秒。我们的面具涵盖更完整的人体。

接不同粒度的视频片段,为准确动作分割,并实现等几个重要的视频理解任务的有前景的结果行动识别、定位和动作模型排名。我们的方法需要最低监管与许多现有方法相比,并显示潜力更复杂的任务,比如人体姿态估计。在未来的工作中,我们计划扩展这个模型和学习更好的人类行为表示动作检测和视频中人体姿态估计。代码为我们的方法和分割结果在三个数据集可从作者的网站。

致谢。这部分工作是由美国国家科学基金会支持的事业拨款(iis - 0845282),美国陆军研究办公室(w911nf - 11 - 1 - 0090)和美国国防部高级研究计划局心眼程序(w911nf - 10 - 2 – 0)

	IOU				Precision				Recall			
	U	U+P	U+P+H	All	U	U+P	U+P+H	All	U	U+P	U+P+H	All
UCF-Sport	40.0	42.1	44.2	**47.6**	67.8	70.4	**73.8**	72.8	54.4	55.2	56.3	**60.6**
J-HMDB	42.7	44.6	45.8	**48.8**	57.2	60.5	**63.0**	62.7	70.0	68.0	67.7	**72.0**
Penn Action	48.9	49.0	49.9	**51.5**	66.1	67.0	69.6	**69.8**	68.1	67.7	65.8	**69.4**

Table 2. Table shows mean IOU (Intersection over union), mean Precision and mean Recall of UCF-Sports, J-HMDB and Penn Action data sets, with only unary potential(U), unary and pairwise potential (U+P), with high order (U+P+H) and our full model (All).

4.3。动作识别和定位

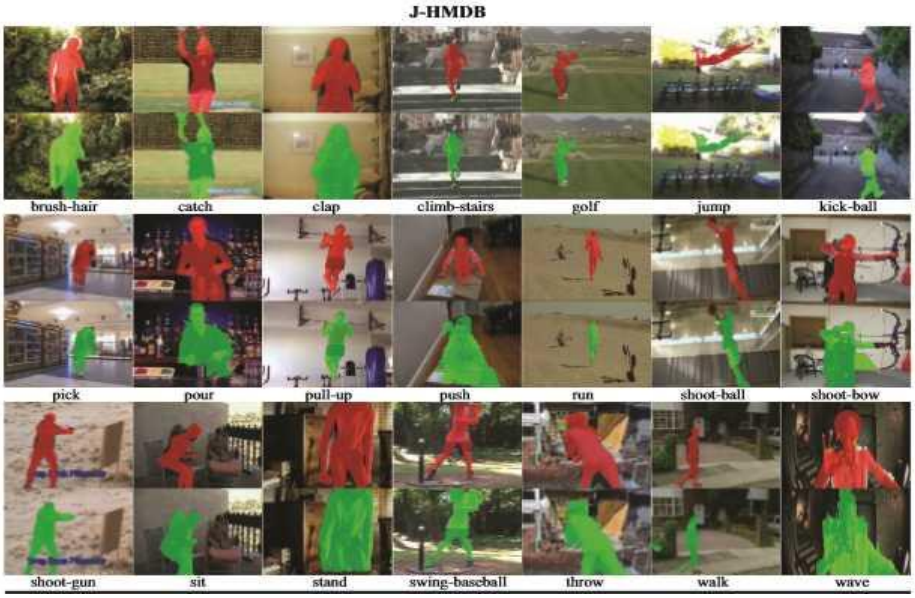
在得到人类行为细分后,我们进一步评估如何影响行动的识别和定位。首先,我们提取密集的轨迹特性在整个视频,2)我们的分割掩模,3)地面真理边界框和 4)地面真理木偶面具 JHMDB(只)。我们发现使用边界框或分割掩模通常可以提高分类精度在整个视频,特别是使用分割掩模通常比边界框得到更好的结果。[16]的发现是一致的评估和证明我们的方法是一个似是而非的工具更好的视频理解。值得注意的是,我们 UCF-Sports 分类精度达到+ 10%的涨幅近有效视频理解论文[24](81.7%)和[14](80.24%)。

此外,在 JHMDB 数据集,我们发现使用分割掩模实现更好的准确性比地面实况木偶面具,也发现在 4.2 秒。我们的面具涵盖更完整的人体。

行动本地化,我们比较平均借据分数[21]和[24],尽管它不是专为定位,我们仍然达到 6–7%意味着借据。

5.结论

在本文中,我们引入一个分层 MRF 模型自动段人类行为边界视频"野生的"。我们做出一些贡献,包括一个强大的人类运动显著特征和一种新的高阶势连



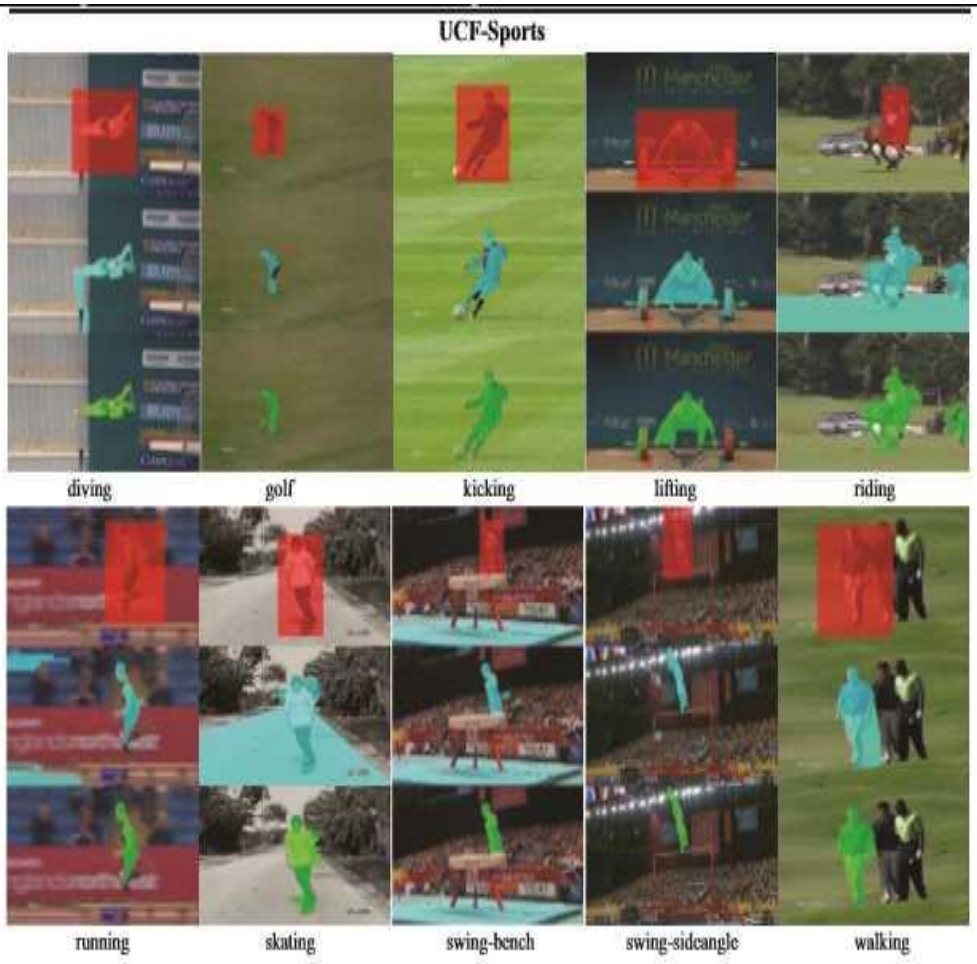


Figure 6. Segmentation visualization of JHMDB and UCF-Sports data set of all the classes. The red is the ground truth mask, the cyan in UCF-Sports is the unary result and the green is our result with full model.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[2] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *European Conference on Computer Vision*. 2010.

[3] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

[4] D. Burschka, J. J. Corso, M. Dewan, W. Lau, M. Li, H. Lin, P. Marayong, N. Ramey, G. D. Hager, B. Hoffman, D. Larkin, and C. Hasser. Navigating Inner Space: 3-D Assistance for Minimally Invasive Surgery. *Robotics and Autonomous System*, 2005.

[5] W. Chen, C. Xiong, R. Xu, and J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[6] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–45, 2010.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

[9] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[10] J. Guo, Z. Li, L.-F. Cheong, and S. Zhou. Video co-segmentation for meaningful action extraction. In *IEEE International Conference on Computer Vision*, Dec 2013.

[11] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2009.

[12] P. Huber. *Robust statistics*. Wiley, New York, 1981.

[13] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *IEEE Conference on*

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *IEEE International Conference on Computer Vision*, 2007.

[19] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, 2011.

[21] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *IEEE International Conference on Computer Vision*, Nov 2011.

[22] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Proceedings of Advance in Neutral Information Processing*, 2010.

[23] J. Leonard, J. How, S. Teller, M. Berger, S. Campbell, G. Fiore, L. Fletcher, E. Frazzoli, A. Huang, and S. Karaman. A perception-driven autonomous urban vehicle. *Journal of Field Robotics*, 25(10):727–774, 2008.

[24] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *IEEE International Conference on Computer Vision*, 2013.

[25] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. In *Journal of Visual Communication and Image Representation*, 1995.

[26] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *European Conference on Computer Vision*, 2014.

[27] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[28] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[29] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *IEEE International Conference on Computer Vision*, 2009.

[30] K. Soomro, A. R. Zamir, and M. Shah. A dataset of 101 human action classes from videos in the wild. Technical re-

[34] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, 2013.

[35] L. Wang, Y. Qiao, and X. Tang. Video action detection with relational dynamic-poselets. In *European Conference on Computer Vision*, 2014.

[36] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[37] C. Xu, S. Whitt, and J. Corso. Flattening supervoxel hierarchies by the uniform entropy slice. In *IEEE Conference on Computer Vision and Pattern Recognition*, Dec 2013.

[38] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *Proceedings of European Conference on Computer Vision*, 2012.

[39] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision*, 2013.