指导教师： 杨涛　　提交时间： 2015.3.28

The task of Digital Image Processing

数字图像处理

School of Computer Science

No: 1

姓名： 麻佳杰

学号： 2012302412

班号： 10011202

带有节段文法的动态视频解析

Hamed Pirsiavash
Deva Ramanan

MIT
UCI
hpirsiav@mit.edu
dramanan@ics.uci.edu

摘要

现实生活中人类活动的视频在不同规模上展示出了暂态结构;长视频是典型地由许多鲜活的例子构成的,并且长视频中每一个例子本身都是由带有不同的持续且顺序的子运动而构成。暂态文法能大概建立出类似于分层结构一样的模型,但是他们很难被计算性地应用于长视频数据流处理。我们描述了一些当利用一台有限状态机器承认推断时能捕捉分层暂态结构的简单文法。这就使得分析成为线性时刻,永久保存而且自然地在线。我们训练我们的文法参量使用一个潜在的结构模型 SVM,在此模型中潜在的运行为自动学习。在共同的、崭新的 50 万框架的多帧数据集连续 YouTube 视频上,我们陈述了我们方法的有效性。

1. 介绍

我们关注在连续的现实生活中视频数据流的行为分类和节段任务。许多早期的行为辨识工作关注的是预先分类。然而,这忽略了一些带有未知数的行为例子的视频加工数据流的复杂性,每一个都带有广泛的持续性和开始/结束时间。更进一步,行为本身呈现除了内在暂态结构。举个例子,一个“茶艺“行为需要很多时间并且要求许多潜在动作的叠加,例如,加热水,准备茶叶,浸泡茶叶,把茶水倒进被子。每一个潜在动作会持续并且有时会有暂时的排序。

我们的工作: 在这个工作里,我们改进了在叠在暂态规模中能回溯分层分析长视频数据流的算法(图 1)。我们的算法是基于一类文法,这些文法能将视频拆分为行为节段,并且递归地将行为拆分为潜在行为节段。我们描述了通过一台有限状态机能被在线分析的专门文法。我们的文法能够线性的测量是视频的长度,还能操作处理有界的记忆,对加工数据流视频的连续性也非常重要。重要的一点,我们描述方法为了从局部标注的数据学习文法。我们假定测试视频是有行为标注的,并且描述一个最大边缘化学习算法给潜在地推测潜在行为结构。

改进: 许多行为辨识的数据集是由预先节段的剪辑构成的。从 YouTube 运动剪辑中,我们已经建立了一个连续暂态数据流数据集。我们 5 小时时长的视频数据集包含了持续的视频数据流带有多种类的行为叠加并附有潜在行为结构。我们将会让这成为可能去让社会支持鼓励我们的更进一步的调查研究。

2. 相关工作

我们建议读者去参考最近的调查[20，1]为了能有一个全面的对行为辨识的大体认知。我们关注的方法大多数都是与我们的方法相关的。

时空模板： 许多行为识别的方法是基于时空模板[11，13，21]被定义域不同的特征包括光流[19,4，24]并且特征分布直方图也是建立于时空“单词”[31,16]。我们用单词代特征作为我们的数据模型。

潜在暂态结构： 被[16,28]所鼓励，我们模拟行为模板作为子运动的成分。举个例子，我们知道一些举重多做应该被模拟为‘抽’子运动紧跟着‘按’子运动（图 1）.然而，过去许多工作仅在单动作视频片段评估这些模型。我们的工作不同之处在于，我们使用语法组成的多个实例，行动起来，以产生全局最优的视频解析。

行为节段： 为了加工带有多行为叠加的长视频数据流，我们必须解决一个暂态时间切割问题。在历史上，这是最常见的利用隐式 Markov 模型（HMM）解决的问题。早起方法可以追溯到有限状态机模型[33,17]，然而许多近代工作已经发现了差异性的变异就像 CRFs[32]。这也存在了大量的文献针对于 HMMs 模型的扩展，在那些文献中声明了可变长度序列的生成；这有时叫做分层 HMMs[5,30,17]

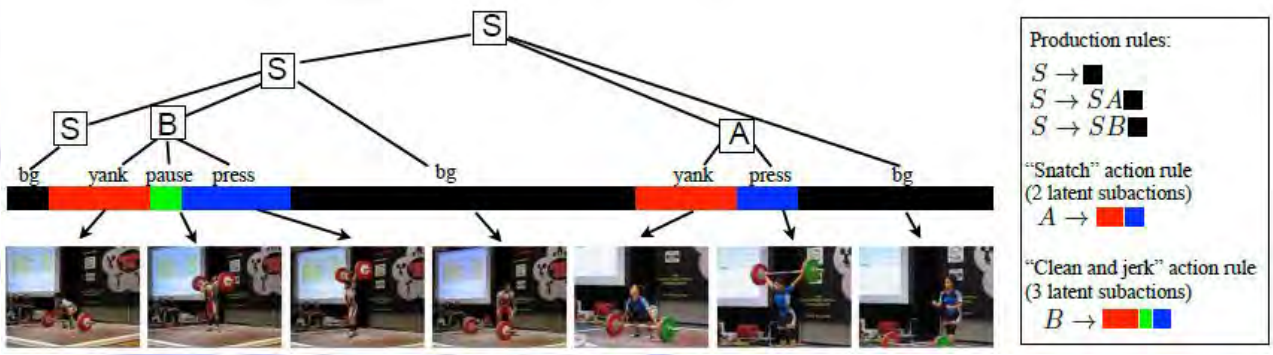

图 1，左侧，我们展现了一个视频的分层分析，将视频切割为动作和子动作。这些动作有非终端符号展示，而子动作和背景切割则是被终端符号展示。我们在右侧展示了文法联系。

或者半 Markov/节段式 HMMs[8,23]。大多数与我们工作相关的是[25,6]，它使用半 Markov 模型去将视频流分割为动作。当模拟地分解动作为子动作时，我们的文法做同样的工作。

文法： 许多早期打的工作，为了姿态和事件的辨识[7,15,22,26]，已经探索了上下文无关算法（CFGs）。属性文法[10]和间隔逻辑[29.12,2]概况 CFGs 包含上下文相关约束，典型地带有更昂贵的推理成本。由于这个负担，许多早期的工作在预先加工阶段将文法应用在稀

疏的原始动作检测环节。这使得带有检测数量的推理规模优于视频的长度。我们描述的这个分层文法能用一个有限状态机模拟出动作和子动作的模型。这便允许我们可以直接使用我们的文法去更有效地加工一个视频的所有框架。

3. 文法模型

我们最主要的贡献是一个动作的分层规律文法，它在 3.4 节被描述。为了使用它，我们为通用 CFGs[14]最先复习了 CYK 分解算法，因为它组成了我们最有效分解算法的基础。一个值得注意的层面是我们的文法是一个能产生多种多样长度分段而不是单独环段；为了使用这个功能，我们最先给 CYK 分解器构建了一个简单的模型去操作这些产生式规则。

3.1 上下文无关文法：

在乔姆斯基范式（CNF）中一个加权的 CFG 是被一下所指定的：

1. V 是一个非终止符有限集
2. Σ 是一个终止符的集合
3. R 是一个规则集由范式 X→YZ 或者 X→w 当 X,Y,Z∈V 并且 w∈Σ 。每一个规则 r∈R 都有一个联合的多元组 s(r,i,k,j)来在边界点 i,j 带有一个过渡点 k 上实体化它。

一个通用的 CFG 包含着规则范式 α →β ，当α 是任意的非终止的并且β 是一个任意的字符串不论是否终止时。我们用 nv 表示非终止符的数量，并且用 nr 表示规则的数量。任何一个 CFG 都可以被改写为 CNF 范式通过添加新的带有“假设”非终止符的规则。

已知的一个序列 w1; : : : ;wN, CYK 算法是一个 O(nRN3)动态规划算法来计算最佳得分解析[14]。在 O（nvN2）规模下，这个算法会计算出一个局部解析的表。CYK 会明确地计算出每个可能节段的最优解析以及这个节段的每一个可能的符号标注。π [X,i,j]是一个被迭代计算的关键值，这个节段最优解析的得分从 i 帧开始，在 j 帧结束，并且被标记为符号 X ∈V。

在 CYK 算法中，我们最先初始化表的“最底部”横坐标，使其代表每个单帧长节段的最优解析：

$$\pi[X,i,i] = \max_{r\in\{X\to w\}} s(r,i) \qquad \text{for} \quad i = 1\ldots N \quad (1)$$

我们现在可以填写“第二层”横坐标，让它代表 2 帧节段的最优解析。对于一个 l

帧的节段，我们可以审查所有可能的 l-1 拆分以及所有可能产生出这个节段的 nr 产生式规则。通过观察整体的已经被计算好的底层数据，每一个节段都能被得分。然后我们可以取最大值，并且更新当前 l 帧长节段的输入。我们正式描述的算法在 Alg.1 并且可视化其核心圈在 Fig.2.。

```
for l = 2 : N do
    for i = 1 : N − l + 1 do
        j = i + l − 1;
        π[X, i, j] =
            max_{r∈{X→YZ}, k∈{i...j−1}} s(r, i, k, j) + π[Y, i, k] + π[Z, k + 1, j];
    end
end
```

Aig.1

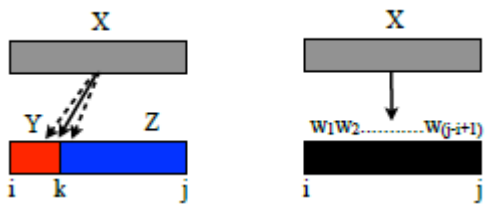

Fig.2

3.2 节段式的上下文无关文法：

在这一节，我们描述一个 CYK-解析的拓展，它能允许产生多叠加终止符的产生式规则。尽管我们的拓展有些直截，我们也没有看到它在文献里派生出来并且使它完整化。我们稍后将会展示节段水平的产生式规则对节段持续性捕获约束是非常重要的。考虑的一个 CNF 规则集被那些带有允许非终结符产生一个 k 帧长终结节段的规则所增强：

$$X \to w_{1:k} \quad \text{where} \quad w_{1:k} = w_1 \dots w_k \qquad (2)$$

每一个上方的规则都有一个联合的多元组 s（X→w,i,j）去搁置一个（k=j-i+1）-元素节段，起始于 i 终止于 j。我们把这种 CFG 成为节段式 CFG（SCFG）。SCFGs 能被当做一种*规则规模数为 N* 的 CFG，并发推理。一个选择性的方法会是将节段长度 k 在属性文法中编码为一个属性。然而，产生式规则的结果约束是上下文相关的，因此它们不能再用 CKY[27]分解。

我们可以显示出，在一个简单的模型里，CYK 可以接受来自表（2）的规则，并且

在不增加任何复杂度的情况下，保持算法在 $O(nRN^3)$ 内。用以下的连个公式来代替 CYK 的公式 4：

$$v = \max_{\substack{r\in\{X\to YZ\}\\ k\in\{i\dots j-1\}}} s(r,i,k,j) + \pi[Y,i,k] + \pi[Z,k+1,j]$$

$$\pi[X,i,j] = \max_{\substack{r\in\{X\to w_{1:k}\}\\ k=(j-i+1)}} \Big(v, s(r,i,j)\Big) \qquad (3)$$

对于表的每一个输入，我们可以像从前一样查询出他的推导（等价于 CYK 的公式 4），但现在我们还能查询来自最底层的节段终止推导。我们现在仅仅需要检查（k=j-i+1）长度的终止推导，在那其中存在的最大长度为 nR。这就意味着 CFG 的每个分割解析都能在 O(nRN3). 复杂度内。

3.3 规则文法：

由于伴随着视频长度，计算与存储的增长变得超线性，因而 CYK 解析可能很难英语用长视频。乔姆斯基阐述了一个上下文无关文法的特殊情况，被大家熟知为有限状态文法或者规则文法[3]，它由限制性规则构成，这个限制性规则带有一个单独非终止符伴随着任意数量的终止符在公式右侧：

$$X \to Yw, \qquad X \to w \qquad (4)$$

这类文法不能模拟出递归定义的语言，例如带有括号嵌套的字符串[14]。然而，它们可以被解析通过有限状态机将他们制造为第一命令状态的 Markov 模型的方法，具体步骤如下：添加一个虚拟的非终结符去将产生式规则转换为 CNF（如果他们没有就绪），定义一个 Markov 模型给每个非终结符，并且最终定义一个从 Y 到 X 的过渡给每一个产生式规则。这样的文法可以被一个标准的维特比解码器解析在复杂度 O（nRN）内，其中 nR(过渡的数量)上界是 n_V^2 [5]。因此规则文法自然地逼稀疏 Markov 过渡矩阵有优势。

规则文法被广泛应用于模式匹配中的指定正则式[14]。我们在图 1 中描述了一个规则文法用来解析举重视频。举个例子，一个挺举动作（B）被定义为一个“字符串”带有，

抽，停，举的终结符。在图 4 中规则文法表现为 CNF 的形式。即便这样的长距离约束能被增强状态的 Markov 模型（相当于虚拟非终结符）和稀疏过渡阵列表述，但正则式的产生式规则仍是一个更简洁的表达。

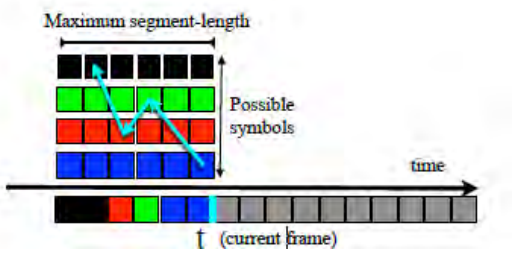

图 3

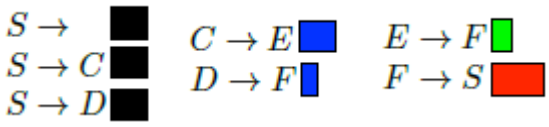

图 4

3.4 节段性规则文法：

我们定义了一个节段性 RGs（SRGs）为一个带有产生式规则的 RG，这些产生式规则能产生出随意长度的连续终端：

$$X \to Y w_{1:k}, \qquad X \to w_{1:k} \qquad (5)$$

SRGs 可以被转化问 RGs 通过添加虚拟非终结符。我们表明可以制造一个解析器来直接处理上述的节段性规则。我们的 SRG 解析器可以被当作一个节段性 CYK 解析器的特例或者一个维特比解码器的拓展。

限制性的 CYK： 假设得分可以被记录为 s(X→Yw,i,j),其中 i 是开始 j 是（k=j-i+1）节段的终结。假定那些节段可以变成 L 个元素长。SRGs 能被解析在复杂度 O（nRNL）里，其中 nR 是被强调的非节段性文法的规则数量。这个算法类似于 CYK 除了我们需要去维持额仅仅第一列表输入π [X,i,j]，当 i=1 时。我们这么写当π [X，j]，从 1-j 的最优解析得分被给予了符号在 j 处是 X。为了简化符号，让我们定义一个控符号{}，它允许我们写（5）的两个规则通过一个简单的范式。初始化

$\pi[\{\},1] = 0$ and $\pi[\{\},j] = -\infty$ for $j > 1$:

1 **for** $j = 1 : N$ **do**

2 $\quad \pi[X,j] = \max_{\substack{k\in\{1\ldots L\} \\ r\in\{X\to Yw\}}} \pi[Y,j-k]+s(r,j-k+1,j)$

3 **end**

算法 2：我们的 SRG 解析器，在图 3 中可视化。X 的最优推导内环搜索从前面的非终止符 Y 过渡，在 j-k 的某些位置（在那之后 k 仍旧是被分割的 X）。在每次迭代里，算法检索超过 L 可能的过渡点并且 nR 可能的规则为了那些过渡点，是整体计算在 O（nRNL）中。

在线解析： 如果我们解释 j 为索引时间，那么算法自然地在线了。 在任意一个点，我们能计算出最优解析的得分知道 X π [X,j]的最大值。在线解析的存储是 O(nVL)-独立于 N-因为我们仅仅需要为过去的 L 个时间节点存储表的输入。这便使得长数据（甚至是多数据流）规模化。我们在图 3 中可视化了我们解析器的运作。

稀疏 Marokv 解码器： 就像 RGs 能被记为状态之间带有稀疏过渡阵列的 Markov 模型一样，SRGs 能被记为节段标志之间的带有稀疏过渡阵列的半 Markov 模型。定义一个 Markov 分割标示对应给每一个非终结符，并且定义一个分值给规则 s(X ! Y $w_{1:k}$; j) ，这便是从分割标示 Y 到分割标示 X 的过渡得分，其中从 j-k+1 到 j 的帧都被 X 标示[23]。SRGs 是有用的因为它们充分利用简洁正则式来给节段之间的长距离约束编码。

解析树： 我们可以通过存储算法第二条公式的最大值指针来回复最有解析树。让 R[X; j] 和 K[X; j]指向最大值规则 r 并且释放 k。我们可以呈现一个带有(r,j)对集的解析树，这个对集有回溯 R 和 K 得到，在最后一帧的最高得分符号处初始化。我们用蓝色箭头展示了这些回溯的最大值指针在图 3 中。我们像以下编码一棵解析树：

$$P = \{(r_m, j_m) : m = 1 \ldots M\}$$

其中，M 代表解析中不同的节段，jm 是第 m 个节段的最后一帧，rm 是规则编号，它的公

式右侧就是节段。

4. 模型结构

我们将所有动作都用下述格式的产生式规则模拟：$A \to xyz$

其中，A 是一个动作，x,y,z 可变长度的子动作、我们选择出最优数量的子动作通过交叉验证，但我们发现三个工作的相当一致。我们的 SRG 文法符合下列格式：

$$S \to b, \quad S \to SAb, \quad S \to SBb, \quad \ldots$$

其中，S 是一个有效的解析，b 是一个可变长的背景终端（混合叠加），并让解析器选择最优的一个。在实际工作中，我们发现了一个单背景终端工作的相当好。那个节段水平产生式规则是非常重要的因为它们能捕捉全球特征在节段像持续性节段下。

已知一个连续数据 D 和一个后补解析 P=$\{(r_m, j_m)\}$

我们可以记录他的得分在我们的加权 SRG 文法下：

$$S(D,P) = \sum_{m=1}^{M} s(D, r_m, i_m, j_m) \qquad (6)$$

其中$i_m = j_{m-1} + 1$并且我们明确地在数据 D 上表示出每每段的得分$s(r_m, i_m, j_m)$

为了更好地描述我们实质的数据模型，让我们定义 Xm 为规则 rm 右侧终止符：

$$s(D, r_m, i_m, j_m) = \alpha_{x_m} \cdot \phi(D, i_m, j_m) + \beta_{r_m} \cdot \psi(k_m) + \gamma_{r_m} \qquad (7)$$

为了描述我们模型的参数，我们将使用一个为举重定义的文法的确切的例子，在图 1 中展示，并且在图 4 中定义。终止符对应子动作，并且抽象符号对应动作。一个挺举动作被定义为一个抽，停，按的子动作，但一个抓动作被定义为一个抽和按。

数据： 第一项是一个数据项，它将节段特征从地 im 帧到 jm 帧提取出来，并用一个调整子动作模型（终端）给它们打分 Xm。我们计算出一个特征代描述符∅给每个分段，其中一个特征是一个时空视觉单词[31,28]。我们能解释一个α_{x_m}为被调整为实际子动作的一个模型；举个例子，$x_m = \text{yank}$的模型将会被调整为时空单词引爆蜷缩的姿势。α 是一

个确切符号 Xm 优于确切规则 rm；这使得不同规则可以为了平等符号分享数据模型。举个例子，挺举和抓动作在图 4 中分享了同样的子动作模型。

暂态模型： 第二项相似于一个时间“先前”其支持实际分割长度 $k = j_m - i_m + 1$ 为一个子动作。β 是确切的规则 rm 并且不是确切的符号 Xm。这便允许一个抽子动作有不同的前提提供给它长度，这取决于那个动作被实例化。特别地，我们定义了 $\psi(k_m) = [k_m \quad k_m^2]$。这意味着参数 β_{r_m} 可以被解释为余下的位置，并且这个弹簧的刚性决定了这个节段的理想长度。在我们的实验结果中，我们展示了这样的暂态项，它们对保证良好的分割精度是非常重要的。更深一步，这样的约束对标准 HMM 编码是非常困难的。

规则模型： 最后一项 γ_{r_m} 是一个“先前”标量或者是支持使用确定规则的基础。举例说的话，这可能编码一个现实：挺举逼抓更可能出现。

5. 学习

我们的模型在（7）中是线性的在模型参数 $w = \{\alpha, \beta, \gamma\}$

这允许我们记录一个解析的得分就像一个线性函数：$S(D, P) = w \cdot \Phi(D, P).$

即便我们在潜在框架里学习了子动作，我们最初描述的全程监控的方案可以简单说明。

全程监控学习： 假定我们被给予了测试中的带有地面实况视频数据的解析 $\{D_n, P_n\}$ 和一个手动指定的产生式规则集Γ 。我们期望为了数据模型α ，赞叹模型β 和规则成分γ 去学习权值 w。权值应该给正确的解析评分高而不正确的解析评分低；我们通过定义一个结构预测学习函数实现它：

$$\arg \min_{w, \xi_n \geq 0} \quad \frac{1}{2} w \cdot w + C \sum_n \xi_n \qquad \text{s.t.} \quad \forall n, \forall H_n \quad (8)$$

$$w \cdot \Phi(D_n, P_n) - w \cdot \Phi(D_n, H_n) \geq \text{loss}(P_n, H_n) - \xi_n$$

上述的线性约束阐述，对于每一个测试视频 Dn，正确的解析 Pn 应该得分超过一个假设的解析 Hn 通过一些值有丢失组（Pn,Hn）给出。我们用不同的丢失函数测试他们。许多方法使用海明损失，这简单的输出了带有不正确符号标签的帧数。我们也考虑了一个更简单的 0-1 损失，它定义了一个后补解析作为非正确如果没有一个属于它的过渡是靠近地面实况的过渡。最终，我们允许这些约束轻松地满足使用松散变量。上述等价于一个结构 SVM，为了它血多更好地解决方法还存在[9]。学习中主要的计算步骤要求解决一个丢失增

强的推断问题，其中对于每一个测试案例，都要找都一个最差违规解析：

$$H_n^* = \max_H w \cdot \Phi(D_n, H) + \text{loss}(P_n, H) \qquad (9)$$

我们的丢失函数可以被（7）的数据项吸收，这就指明我们高效的解析器 SRG 能被应用于寻找这样一个违规解析。

子动作的潜在学习： 在学习中明确一棵满的解析树是非常昂贵的或者含糊的。在我们的场景我，我们在不给任何子动作标签的情况下，被给予了每帧都带有动作标签的视频流。我们给视频 n 做一个动作标签 An。在这个集合里，我们用一个潜在的结构 SVM[34]自动地评估子动作。我们通过迭代下列步骤使用 CCCP 算法[35] ：

1.给定的 O 型 w，推断一个解析 Fn，赋予每一个视频 n 一致的 An。

2.给一个满解析集合{Pn}，通过解决 QP（8）学习一个模型 w，伴随着一个动作明确损失。

第一步是用一个无限损失通过修改损失函数（9）去处罚一个不一致的解析。这个损失也可以被数据项吸收。在第二步中，我们定义了一个动作确切损失（An,H）,它处罚不一致的动作过渡而不是子动作过渡。即使数据模型，暂态先前，和规则赋权都已经被充分的学习，我们还必须明确产生式规则。另外，我们还必须明确一个开始状态给所有迭代。两者都在下面阐述。

始化化： 我们的文法可以给动作类之间编码。然而，在我们的实验中，我们分析那些饱饭单独动作类实例的视频。简单地说，我们学习分离文法，这样每个动作类都使用下述的通用规则集：$\{S \to b, S \to SAb, A \to xyz\}$.

我们通过分割每一个动作实例 A（在测试集中）为一个混合的等大小的子动作 xyz 来初始化迭代。在每一个动作集里我们使用交叉验证法去选择最优子动作数，即使我们找的了 3 个工作相当的一致。最后一个推断的子动作标签可能相当的非制式化（潜在地评价测试和测试数据），就像图 6 和 7。我们潜在学习的策略跟工作[28]非常的相似，除此之外我们为了全面解析一个带有叠加动作而不是单独分离动作的视频流而学习规则。

图 5

6. 实验

数据缺失： 通俗动作的基准就像 TRECVID 中的 MED 挑战一样，由带有单动作的短视频剪辑构成，因此不适于评估我们的方法。先前动作分割方法经常被人工构建视频评测，通过被连接在一起的预分割视频剪辑获得[6,25]。这是个清楚的限制。而且，我们不清楚一些由连续连续非脚本的人类活动构成的基准视频数据集，在监控录像[1]和耐受相机记录之外[18]。

连续奥运数据集： 为了解决这方面的不足，在[16]的鼓舞下我们已经收集了我们自己的数据，它介绍了一个业余的现实的数据集（但已预分割）由 16 个不同奥运运动的 YouTube 视频剪辑构成。从 YouTube 上我们已经收集了由一个 8 动作集构成的连续视频。我们的连选奥运数据集包含了几乎 5 小时或者 50 万帧在这个现实视频镜头上。每个视频包含了一个平均大约 6 个的动作实例，每个实例都被开始/结束帧符号所注解。图 5 阐述了一个每个视频的示例帧。我们将释放这个数据集来激发我们更深层次的现实世界动作解析的研究。在我们的实验里，我们使用一个 50-50 训练/测试分裂。

进化： 我们假设我们被给予了一个已知动作类的视频，然后运用我们的动作明确文法来解析视频为动作和子动作。因为我们没有地面实况子动作标签，我们仅仅改进动作标签的精确度。我们在图 6 和 7 中展示了定量的结果，并且包含了我们补充的解析视频材料。我们建议读者去配图里面找细节分析。我们被给予了一个后补连续解析，而且我们用两个方法定量的分析它（图 8）。

每一帧： 我们计算帧的数量并匹配后补和地面实况的动作标签。 **动作监控：** 就像 PASCAL 改进的物体监控，我们认为我们的解析器能返回后补动作分割监控的值。一个监控

应是真实积极的，如果如果它与地面实况动作段之间的重叠（并集/交集）大于预定阈值。我们使用 40%作为默认值，就像我们发现它与真实目测一致。我们也改进了其他的阈值。

基准线：我们看到了表现不佳的简单基线，如扫描窗口模板和单帧 HMM。我们对[28]的最先进的模型进行比对，它学习了一个动作明确地**子动作**模型柏油 3 个子动作，动作被半 Markov 模型所模拟，其中状态对应着潜在子动作。然而，如何使用这个模型去监测单视频中叠加动作尚不确切。我们应用它去浏览窗函数，并且用 NMS 产出一系列为超纲的动作分割。这产生出了可靠地帧标签（36%正确），但不太适应去节段监（2%AP）。这暗指着节段监测的困难性。我么也比较了[25]的模型，它它使用了一个半 Markov 模型去加工整个视频，其中状态对应动作/背景标签伴随着早前超过的暂时持续。这样的模型在[6]中被探索，当被随机函数测试时。算法上地，在我们的文法里这样的模型对应一个不带有子动作的“一阶”版本。它们直接产生一个全面的解析不通过要求 NMS，但却不能对子动作负责。这给 15%AP 增进了精确度。

我们的模型： 我们最终的模型可以被视为一个[28]和[25]的结合，它使用了一个增强的状态空间(被我们的产生式规则定义)去解析动作和子动作。我们最终的精确度是 22%AP。因为我们的解析器是在算法上与半 Markov 模型等价的带有稀疏过渡阵列的，更重要的是它和基准线一样快。我们也将我们模型的 一个版本比作一个无长度早期的持续性子动作。这样剧烈的衰落表现为从 22%到 8%。最终，我们使用一个节段的 CYK 解析器（3）来解析我们的 SRG，它产生出了理想的结果（和预期一样）。然而，我们的接续器大概使用.7 毫秒/帧（像我们的基准线），而 CYK 却大约使用 2.3 秒/帧（比我们慢 1000 多倍）。CYK 能解决更多的通用 CFGs。

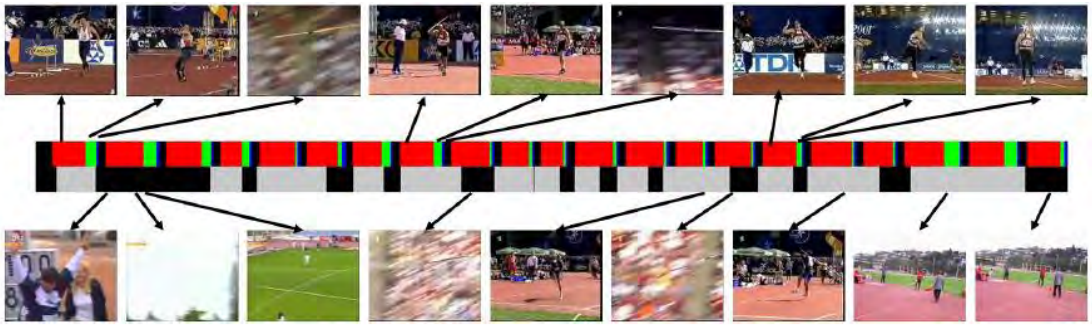

图 6

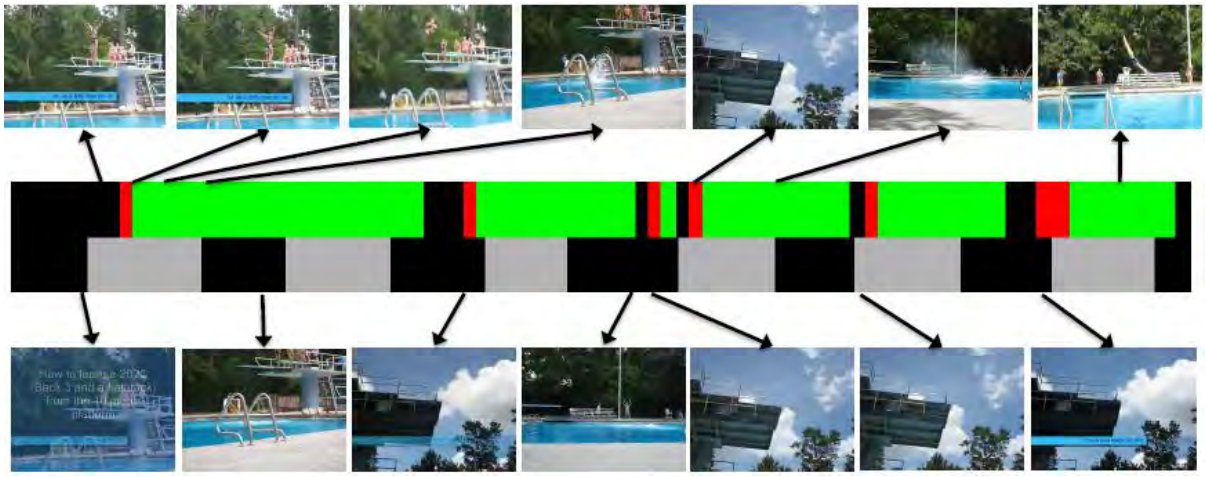

图 7

分析： 我们的子动作模型表现得最好并且带有清晰的结构，例如举重，跳马和跳水。有时候有时次级操作可以在语义上可解释的，但是这不是必须的情况下作为该结构是潜在推断。毕竟，我们的模型在标签帧（62%精确度）是可靠地成功的，但仍旧发现了节段监测挑战（22%AP）。降低重叠阈值（标记一个动作段为正确）从 40%到 10%，提高 AP 从 22%到 43%。行动边界地面实况标签可以是不明确的，所以较低的重叠阈值是合理的评价。更多的是，低阈值仍旧会给能力评分去计算一个视频中的动作实例数量。我们的结果建议我们的解析器能被相当精确地适用于动作实例计数。

总结：我们已经描述了文法段扩展动作分析，并着眼于有效的节段性规则文法。我们展示这样的模型捕获暂态越是在叠加规模中，两者都是在动作和子动作之间。我们介绍了线性的，永久记忆的和在线的并且非常适用于长视频的解析算法。我们同样介绍了最大利润算法来推断潜在子动作构成分布标记的测试数据。为了阐述我们的方法，我们引入了一个新的连续动作的数据集，并表现出令人鼓舞的成绩，它超过了一些标准的基准方法。

致谢：NSF 资助 0954083，ONR，穆里格兰特 N00014-10-1-0933，和英特尔科学与技术中心 - 视觉计算资助了这项研究。

Segment Detection (average precision)/ Frame Labeling (% frames)

Action name	Subactions [28]	our model (no length prior)	Segmental actions [25]	our model
weightlifting	0.20 / 0.51	0.13 / 0.74	0.27 / 0.6	0.53 / 0.6
javelin	0.006 / 0.54	0.11 / 0.64	0.36 / 0.59	0.36 / 0.65
long-jump	0 / 0.26	0.02 / 0.46	0.10 / 0.54	0.10 / 0.71
vault	0 / 0.28	0.12 / 0.49	0.06 / 0.34	0.11 / 0.69
bowling	0.007 / 0.39	0.08 / 0.45	0.21 / 0.51	0.25 / 0.54
diving	0 / 0.48	0.10 / 0.53	0.02 / 0.45	0.08 / 0.62
hammer-throw	0.006 / 0.25	0.04 / 0.36	0.16 / 0.29	0.23 / 0.63
tennis	0 / 0.15	0.01 / 0.37	0 / 0.37	0.08 / 0.52
Average	**0.027 / 0.36**	**0.076 / 0.51**	**0.15 / 0.46**	**0.22 / 0.62**

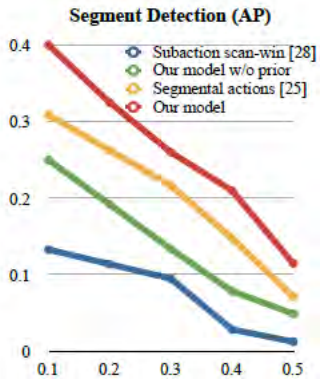

图 8

参考

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16, 2011.

[2] W. Brendel, A. Fern, and S. Todorovic. Probabilistic event logic for interval-based event recognition. In *CVPR*, 2011.

[3] N. Chomsky. Three models for the description of language. *Information Theory, IRE Transactions on*, 1956.

[4] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *CVPR*, 2003.

[5] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

[6] M. Hoai, Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011.

[7] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE PAMI*, 2000.

[8] J. Janssen and N. Limnios. *Semi-Markov models and applications*. Kluwer Academic Publishers, 1999.

[9] T. Joachims, T. Finley, and C. Yu. Cutting plane training of structural SVMs. *Machine Learning*, 2009.

[10] S. Joo and R. Chellappa. Attribute grammar-based event recognition and anomaly detection. In *CVPR-W*, 2006.

[11] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, pages 1–8. IEEE, 2007.

[12] S. Kwak, B. Han, and J. Han. On-line video event detection by constraint flow. *IEEE PAMI*, 2013.

[13] I. Laptev and P. Perez. Retrieving actions in movies. In *International Conference on Computer Vision*, 2007.

[14] C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[15] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *National Conf on Artificial Intelligence*, 2002.

[16] J. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. *ECCV*, pages 392–405, 2010.

[17] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *CVIU*, 96(2):163–180, 2004.

[18] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[19] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *IJCV*, 23(3):261–282, 1997.

[20] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 2010.

[21] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008.

[22] M. S. Ryoo and J. K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *IJCV*, 82(1):1–24, 2009.

[23] S. Sarawagi and W. Cohen. Semi-markov conditional random fields for information extraction. *NIPS*, 2004.

[24] E. Shechtman and M. Irani. Space-time behavior based correlation. In *IEEE PAMI*, 2007.

[25] Q. Shi, L. Cheng, L. Wang, and A. Smola. Human action segmentation and recognition using discriminative semi-markov models. *IJCV*, pages 1–11, 2010.

[26] Z. Si, M. Pei, B. Yao, and S. Zhu. Unsupervised learning of event and-or grammar and semantics from video. ICCV, 2011.

[27] K. Slonneger and B. L. Kurtz. *Formal syntax and semantics of programming languages*. Addison-Wesley, 1995.

[28] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[29] S. Tran and L. Davis. Event modeling and recognition using markov logic networks. *ECCV*, 2008.

[30] T. T. Truyen, D. Q. Phung, H. H. Bui, and S. Venkatesh. Hierarchical semi-markov conditional random fields for recursive sequential data. In *NIPS*, pages 1657–1664, 2008.

[31] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[32] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*. IEEE, 2006.

[33] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *PAMI*, 21(9):884–900, 1999.

[34] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, 2009.

[35] A. L. Yuille, A. Rangarajan, and A. Yuille. The concave-convex procedure (cccp). *NIPS*, 2:1033–1040, 2002.

10011202

2012302412

麻佳杰

2015.3.25