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：李治

学号：2012302534

班号：10011207

对于视频事件检测的时间序列建模

Yu Cheng, Quanfu Fan, Sharath Pankanti
IBM T.J. Watson Research Center
{yucheng,qfan,sharath}@us.ibm.com

Alok Choudhary EECS Department, Northwestern University
choudhar@eecs.northwestern.edu

摘要

我们提出了一种对于视频事件检测的新方法即通过时间序列建模。利用现有的许多关于视频分析的核心方法（比如动作，活动和时间识别等）。不同于之前的作品，在语义事件级别下做时间建模，我们建议在子事件级别下，不使用事件注解，来对数据进行时间依赖的模拟。这就使我们的模型从基础事实中摆脱出来，同时解决了前期工作在时间上建模的一些限制。基于这种想法，我们通过从视频中获得的一序列的视觉词来表示一个视频，并应用这些序列Memoizer[21]在视觉序列的一个时间上下文下中捕获远距离的依赖。对于一个视频事件在联合执行分割和分类时，此数据驱动时间模型与事件分级进一步整合了。我们证明了我们在两个具有挑战性的数据集的视觉识别方法是有效的。

1. 简介

由于视频内容的指数级增长，因此极具需要能够实现智能视频的分析和理解。其中，视频事件检测在许多的应用中起着重要作用（扮演着重要角色），例如监视，主题发现和内容检索。事件检测的任务包括确定一个视频中某一事件的时间范围（即 **when**）和它的位置（即 **where**）。如今虽然有越来越多的努力来解决这个问题，但由于混叠问题，使它仍然是相当具有挑战性的，例如事件大量的内部差异，事件不同的持续时间以及难以避免的背景杂波等因素的干扰。

在这项工作中，我们通过利用事件中的时序依赖关系，来解决我们进行视频事件检测处理的问题。逼真的（真实的）视频事件往往是非独立的（具有依赖性），由于场景的原因，展现出或多或少的相关性。正如图一所示，在一个机场监控的环境下，**PeopleMeet**（一乘客靠近服务台寻求方向信息）依据指向（工作人员指向的方向）的信息行走并通过了 **SplitUp**（两个人分裂向上）。在机场场景中关于事件更丰富的时空格局都由图 **1** 可见。

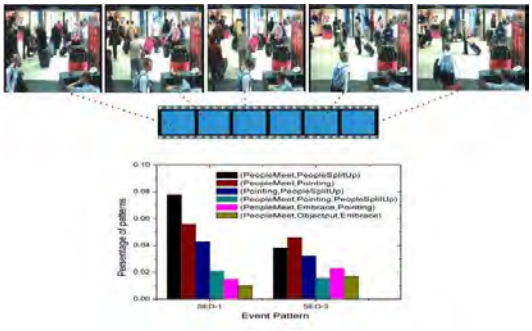

图 1.时序模式在 SED 数据[15]中展出了两摄像头。顶部图像示出这样一个例子：PeopleMeet→Pointing→SplitUp（在变焦和颜色下的最佳视图）。

时间关系模型和上述的结构已经成为人类行为，人类活动以及事件识别的关键所在。动作识别的方法通常集中

于捕捉动作的基本时间结构（即内部的依赖关系），无论是通过特征的表示[11]或使用更复杂的模型[13，8，23]。与此同时，活动识别的工作主要就是尝试去探索一个活动中原始动作之间的时间关系（即内部的依赖关系）通过图形模型，例如 HMM 和 DBNs 模型。

虽然时间建模已经在视频的理解上得到了巨大的成功，但是在视频事件的有效分析方面仍然存有问题有待解决。首先时间建模通行是基于一阶马尔可夫即假设只能捕获当前和以前状态之间的一个短暂的相互作用。虽然这可能足以为那些一个活动中动作的时间顺序是很明确和严格区别的人类活动进行建模，但是当应用到探索那些相对松散，有时相对较远距离的经常出现在事件数据的时间环境（通常未知）下时，它却面临相当大的限制。这 n-gram 模型例如[1]在语音识别方面可能有帮助，但是在实践中，他们经常遭受由于训练数据不足而难以捕获复杂的关系和计算可扩展性问题。其次，在许多的情况下，在一个视频中却只有几个感兴趣的事件并且同时它们常常伴随有空事件和背景杂波等干扰。因此，马尔科夫假设将严重偏向于空事件以及消弱真实事件之间那些导致表现欠佳的依赖关系。第三，大多数方法从实际情况直接建立时间模型。这种模型不能发现与未加注释的不顾依赖性有多强相关的时间模式。例如，如果任何两个事件在注释中不可用的话，在上面的例子中 PeopleMet，Pointing and SplitUp 之间的关系将不能被捕获。

为了解决上述的限制，对于视频事件的检测我们提出一种基于事件建模的新方法。我们制的检测任务作为一个序列建模问题，即我们的目标就是要打破视觉序列成为不同长度的片段同时用一些感兴趣的事件和空事件来标记它们。基于上述的制定，我们首先利用在一种无监督的方式下使用均值聚类算法(Fig. 3)从我们的数据中学习到的一系列视觉词来表示一个视频。然后，我们将应用 Sequence Memoizer (SM) [21]去探索序列中视觉词之间的依赖关系。关于 SM，它是一种最初的无参数的语言建模方法，可以有效的利用离散数据序列模拟远距离的环境内容以及显示在各种各样问题中的幂律特性[24]。更具体地说，SM-基于序列模型具有能够根据一个序列中先前所观察到的内容环境来预测出现什么后续视觉词的能力。正是这种能力使得时间建模成为一种有效的途径，而不严重的依赖于注释。最后，我们将序列模型和事件分类器整合成一个能够共同在一个视频中进行分割和分类的框架。最佳的分割可以通过动态规划来快速的找到，类似于[6]的工作。

我们方法的概述示于图 2. 尽我们的所知，这是极少数方法中的一种方法，即对于一个视觉问题可通过应用一个可行的统计方法去模拟远距离语境下的依赖关系。它呈现出了以前的作品的几个优点。序列模型是建立在视觉词（子事件）之上的，而不是注释事件之上，因此它不需要基础事实。当此出现之后，这种在子事件水平上的时间建模就优于在其本事件水平上的配对。此外，我们的方法自动发现事件背景环境和数据中固有的结构以及利用它们增强检测。我们验证了我们的方法并证明了使用在两个具有挑战的视觉识别数据集的效果。

2. 相关工作

很多关于人类动作识别的方案已经被提出来了。它们中的大多数都是利用事件信息或是特征的表现[11, 20]或是更复杂的模型[13, 8, 19, 23]在预分割剪辑上执行分类。例如，拉普捷夫等。[11]采用时空时间的兴趣点在真实的视频设置中对人体的运动进行分类。Tran & Sorokin[20]开发运动下的功能区学习邻近的指标在 YouTube 视频中

的行为进行分类，Niebles 等。[13]来发了一种无监督模式基于概率潜在语义分析人类行为的检测。更近的，Tan 等人。[19]开发 HMM 模型的一个变体是受训一个最大利润的框架来自动发现视频的歧视性和有趣的部分。Zhang 等人。[23]提出了一种方法，它可以识别

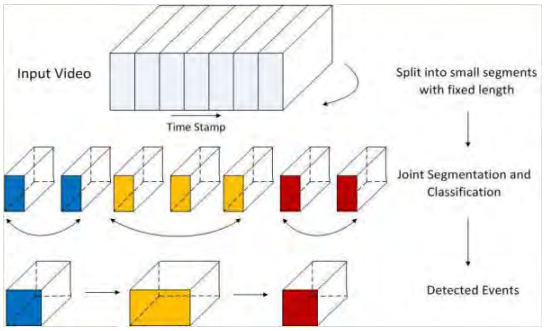

图 2.给定以输入视频，我们的方法是将它均匀的分成以序列事件片段，然后构建该序列顶部上时域模型来在视觉数据序列中捕捉上下文依赖关系。那么这种方法融合了时空模型和事件分级，共同执行事件分割和分类。

本地和远距离运动之间的相互作用，更有效地处理长期活动。

对于那些需要将视一段视频中的动作或时间识别出来的应用程序，滑动窗是一种流行的技术用来讲一个分类器变成一种监测方法。例如，Chen 等人。[3]基于一个费舍尔矢量编码表示及那里了一个事件分类器，用来监控事件并结合多尺度的检测和滑动窗口技术和非最大抑制来进行事件检测。这种方法的难度是在于真正的事件分类阈值的确定。由于这种限制，最近的努力提出去学习用来在较长的视频序列[18，14，6]中同时分割和识别的框架。例如，Oh 等人。[14]开发的线性动力系统来模拟蜜蜂的行为。[6]的工作训练了一个判别性的识别模型，它拥有一个多级 SVM 能够将多级之间的分离冗余最大化，在一个[18]的相似精神中能够使整体的分类分数最大化。

我们工作的另一个方向就是人类行为的识别。行为识别的大部分工作就是通过使用图形化的模型比如隐藏式马尔克夫模型和 DBNS[7，12，17，5]来研究一个活动中原始动作之间的时间联系（即相互依存关系）。然而，这种方法通常需要领域知识用来建立或引导时间建模。

我们的做法与之前的一些方法有以下几方面的不同：1）我们的模型可以通过视频片段的时间关系进行明确的建模，来同时捕获内在依赖型和相互依存关系；2）该模型可以捕获和利用数据中远距离的时序依赖关系；3）模型的构建不依赖于事件注释或基础事实。

3. 我们的方法

不像大多数我们之前关于事件检测的工作，比如说[3]它将视频的分割和事件的分类分开处理，我们的方法是将视频分割和时间分类连同起来进行处理和一个将在第四部分[4]中描述的时间模型。事件模型背后的动机是利用丰富的事件结构和往往存在于事件中以提高检测的依赖关系（相关性）。

3.1 视频表示

给定一个输入视频 X，我们首先将其分成 n 个固定长度 lseg 的时间片段，X = {x 1 ,x 2 ,⋯ ,x n }.然后我们为在动作 SIFT 关键字[2]之上的每段计算词袋的特点。通过使用 k 均值该段被进一步聚类成 k 视觉词，并且每段被分配一个视觉词。最后，该视频是通过视觉词的序列表示 W = {w 1 ,w 2 ,⋯ ,w n }.在我们的实验中，相对于视频的总长度将 lseg 设为不变的，k 通常根据数据的复杂性取值为 600 至 900 范围内。

图 3.说明了一些从数据中学到的子序列。一个直接的观察结果是相同的事件趋向于产生相似的视觉字。一个事件中的一个视觉词可能与另一个来自于不同事件的视觉词进行统计地交互，尽管这两个词暂时很遥远。我们将在第

4 节展示如何对这种远距离的相互作用进行建模。

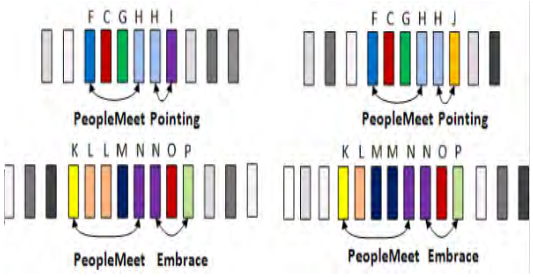

图 3.样本涉及到的事件涉及厥词列。同一时间容易产生相同的视觉单词。为了清晰，我们将空事件的字跳过。

3.2. 联合分割和分类

如上述的视频表示，我们的目标是分区 X = {x 1 ,x 2 ,⋯ ,x n }成 m 个单元，并用感兴趣的事件或者空事件进行标记（Fig. 2）。这个单元是 X 的一组连续片段。令 S={S 1，S2，...，SM}是这样一个分区，其中一个单元 si = Xt1 :t2 = [xt1 , . . . , xt2]和 t1，t2 制定在 si 段的开始和结束的索引。此外，Y ={y 1 ,y 2 ,⋯ ,y m }其中 y i ∈ Y 是分配给 si 的事件分类标签。为了模拟数据中时间环境，我们用一个可视化序列关联 S。分区 S 相对于事件分类的质量可通过

$$f(S,Y)=\sum_{i=1}^{m}\phi(y_i|s_i)+\mu\sum_{1\le k\le i-1}^{i=1}p(z_i|z_{i-k},\cdots,z_{i-1}) \quad (1)$$

进行评估。其中 μ 是从经验数据中学到的折中参数。注意 Z 可以是在 S 的顶部上创建的任何可视数据序列。例如，视觉事件或视觉词的序列。我们将在 4.2 节进一步解释。

这第一项 φ(y i |s i)如式[1]，测量的是单元 S i 是事件 Y I 的可能性。我们为了这个项目在事件 yi 上使用 SVM 分类评分（见 Section5 查看详细）。第二项 p(z i |z i−k ,⋯ ,z i−1)是有我们在 4.2 节中讨论序列模型提供给的。说的简单些，它是在看过之前符号 k 从 zi-k 到 zi-1 后预测 zi 作为下一个字符的概率。当 k=1，该项目降低到充分研究的一阶马尔科夫性。另一方面，如果 k= i- 1，它考虑了这序列的整个历史过程。在以

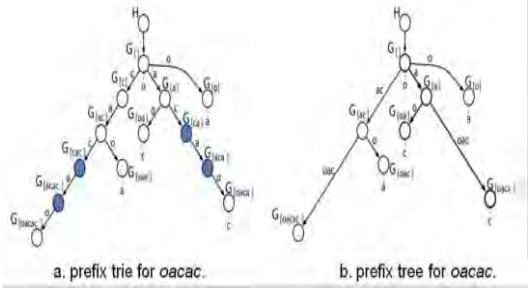

图 4.一个前缀树[21]的例子和一个字符串 oacac 的前缀树的例子

前的工作[22]中，解决这些长期的背景环境是一个有效的关注。然而，一个最近开发的概率模型打破了这种限制，通过在一个离散数据序列中研究背景环境的一个无线长度。

在下一节使用[21]技术详述如何建立时间模型之前，我们简要的描述如何通过动态规划有效地解决上述目标功能（函数）。

3.3. 动态规划

在一个新是视频序列 X 上进行细分可以被浇铸成求一个目标函数 f(s,y)最大化的任务。给定任何视频片 X 0:u 其中 u∈ (0,n].让我们根据 u 考虑一个变化的目标函数 f(S,Y,u)。让 Z u−l:u 作为 X u−l:u 的视觉序列，同时 Z 0:u−l = {z 1 ,z 2 ,⋯}作为之后视觉序列的一个联合。过度函数 θ(u,l)可以被表示如：

$$\theta(u,l)=\max_{l_{min}\le l\le l_{max}}(\phi(y|X_{u-l:u})+\mu P(Z_{u-l:u}|Z_{0:u-l})) \quad (2)$$

其中 y 作为所有可能事件的标签。[l min ,l max]给出一个事件的最大和最小的持续时间，这可以从事实情况中获得到，最后的任务是计算 f(S,Y,len(X))，其可以这样做：

$$f(S,Y,u)=\arg\max_{l_{min}\le l\le l_{max}}\{\theta(u,l)+f(S,Y,u-l)\} \quad (3)$$

不同于[6]使用动态规划彻底的在每一帧中进行搜索，我们的方法只在片段中搜索。对于X市场分析上的实施复杂性是：O(m (lmax −lmin +1) len(X)). Lseg.

4. 时空建模通过序列 Memoizer

为了解决方程 1，我们需要计算一个可视化的标签 zi 基于一个可观察序列的{z i−k ,⋯ ,z i−1 }的概率。在这里为了这个目的我们采用序列 Memoizer（SM）[21]。

4.1. 序列 Memoizer（SM）

序列 Memoizer(A Stochastic Memoizer for SequenceData)是离散序列的一个深不见底的，层次化的贝叶斯非参模型。相对于序列建模的其它技术，SM 能够在一灵活长度的离散序列下有效地学习一个联合分布并且捕捉远距离的依赖关系。这种方法已经证明为一个国家最先进的语言模型和数据压缩。

给一个离散型随机变量 x1 和一个任意长度的 T 的序列 T ={x 1 ,x 2 ,⋯ ,x T }，每次在一个符号集中取值。通过 SM 对序列的联合分布估计是：

$$p(x_{1:T})=\prod_{i=1}^{T} p(x_i|x_{1:i-1}) \quad (4)$$

这暗示每个 xi 预测给出了前面的所有变量 x 1:i−1 的上下文环境。注意，这是从第 n 歌马尔科夫假设作为 T，这可无穷的理论下去。

SM 通过一个前缀特里来表示一个序列(Fig.4.a)，或更有效地前缀数(Fig. 4.b)，它可以从线性时间和空间复杂度的输入字符串构成。根据这种表示，SM 放置一个 Pitman-Yor prior (PYP)来近似树上的每个字序列的频率。这很好解决利用 SM 了训练数据不足的问题，即经常遇到的基于 n 次马尔科夫假设传统序列建模。数学上，s 的概率 s ∈P 给出了它之前的上下文环境 s′，G[s]，可表示为：

$$G[s]|d[s],c[s],G[s']\sim PY(d[s],c[s],G[s']), \quad (5)$$

其中 c 和 d 在皮特曼-尤尔的参数之前。[s] = [ss′]。如图[21]，使用一些特殊的边缘化分析技术，G [s]可以被在线性时间中高效的计算。我么建议读者可以参考[21]来进一步的理解。

在 SM 中，有上下文较后的符号比预测后续符号更重要。基于这种想法，如图 3 例子中所示，两个显示在底部的子序列之间的相似性（即 ...KLLMNNOP.....和... KLMMNNOP）很高，因为它们有着很长后缀。另一方面，由于 KL 在两个序列中都出现了，所以对于预测 L，K 比其它的都显得重要。

4.2. 时间序列模型

一个自然的想法就是将 SM 应用到对一个事件序列进行建模，类似于的 HMM 一样。这很直接并且很容易的做到通过将 Eq.[1]中的 zi 设置为 yi. 我们把这种方法称之为事件级序列模型（ESM）。然而，苏哈的方法，虽然被广泛采用，但要求对基础事实的模型学习。如先前在第一节指出的，这个模型不能够充分的利用 SM，由于事件的稀疏和极不平衡的分布，同时在处理空事件时也不可靠。

采用灵活的视觉序列来实现这公式 1，我们利用公式 1 中的 SM 来建立视觉字序列模型。这样一个在粒度级别的模型，在这里被称之为段级序列建模（SSM），在我们实验室被证明为有效并且相当稳定。这主要在于 a) 相对于那些通常很少的视觉词的真实事件，大量的视觉词更有可能出现在一个幂律分布下；b)序列模型是在纯粹的数据驱动方式下构建的，而不是来自于该事件的注释。

现在我们展示如何 SM 计算 p(z i |z 1 ···z i−1)。记得视觉标签 z 和一个单元 s i = [x t 1i ,...,x t2i]相关联，它可由以序列视觉词[w t 1i ,...,w t2i]（见第三节）来代表。通过考虑到这一点，并应用链式法则，我们可以得到，

$$p(z_i|z_{i-k}\cdots z_{i-1}) = p(w_{t_1},\ldots,w_{t_2}|w_{t_1},\ldots,w_{t_2})_{i-k} = \prod_{j=t_{1i}}^{t_{2i}} p(w_j|w_{t_1},\ldots,w_{j-1})_{j-k} \quad (6)$$

通过在上述式子中设置 k=i-1，最后一项将变成 p(w j |w 1 ,...,w j−1)，这可通过 SM 高效的计算出来。

5. 事件分级

在方程 1，我们需要对包括空事件的所有事件一个单元 si 中所有可能的长度 l(l min ⩽ l ⩽ l max)进行评估。我们注意到事实事件的事的时间长度可以显著的区出来。例如，一个 PersonRuns 事件的最大长度可达 1000 帧，而最小长度仅为 10 帧。如果我们学会用一个固定的时间尺度分类模型，这种多样性将持续带来信息的丢失。因此，我们建议通过第三节中所描述的一个固定帧长度 l 在多时间尺度上学习分类来对原始的视频分割进行匹配。

令 h = (l max − l min)/l seg。如果我们在 l seg 到 h ∗ l seg 每一个维度上为每一个事件训练 h 分类器，那么它足以解决方程 1.早我们的实验中，我们为每个事件使用相同的时间间隔（30-120 帧）并且为每个事件在 30，60，90，120 帧处分别内置 4 个分类器以便能够高效进行。请注意，这与事实相一致：l = h*l seg ,(h = 2,4,6,8),l seg = 15 ,l min ⩽ l ⩽ l max .我们使用多级 SVM[4]对每

6. 实验结果

6.1. 实验装置

我们在 SM 技术的基础上建立了三个序列模型。第一个(ESM-∞)，如第 4.2 节所描述的，在事件水平上进行序列建模，同时考虑到上下文的全长。第二个(ESM-1)，是唯一具有被认为是一阶依

图 5.典型的 SED 数据集的截图。从左至右依次为：Pointing, CellToEar 和 PersonRuns 事件。

赖关系的第一个特殊情况，相类似于 HMM 原理。最后一个（SSM-∞）是我们提出，即一个段级序列模型利用上下文环境的全长。这些模型被集成到第 3 节进行事件检测所描的框架中。

基线 我们实现了由 Hoai et. Al[6]所提出的方法，并利用它在我们评估中中作为主基线。这种方法进行联合分割和人体行为的分类，其基于前两件事件分类分数的边缘最大化。然而，它并不考虑事件之间的时间关系。在 SED 数据集中，除了 HOAI 的工作，陈等人开发的方法。 [3]也被列入到我们比较的范围内。虽然陈的方法通过滑动窗口来进行事件检测，在 SED 数据集上已经达到世界先进水平的性能，在 2012TRECVID SED 排名中已经在 4 事件出 7 种为名榜首。

分类的特点 对于好莱坞事件上的事件分类，我们使用 STIP（科技创新政策）的功能，和 SED 上时空费舍尔矢量功能[3]。我们采用多类 SVM 方法[4]来为每一个事件（动作）训练一个分类器，并为每一个除了陈之外的方法加上一个空事件，它的确通过使用一对多的方法达到多分类的效果。对每一类，4 个分类器分别建在 30，60，90，120 不同时间尺度的帧上。

事件检测与评估 为了产生用于训练我

们模型 SSM-∞的视觉学列，我们使用 k 均值进行聚类一个序列实现均匀化的划分片段。在好莱坞中，在所有测试中 k 被设置成固定的值 200。在 SED 上，我们凭借经验为每一个摄像机确定 k

Events	Hoai[6]		*ESM1*		*ESM-∞*		*SSM-∞*	
	P	R	P	R	P	R	P	R
AnswerPhone	0.64	0.35	0.64	0.35	0.62	0.31	0.67	0.35
HugPerson	0.46	0.37	0.45	0.33	0.44	0.35	0.47	0.37
Kiss	0.44	0.49	0.43	0.51	0.43	0.49	0.44	0.49
SitDown	0.36	0.40	0.37	0.43	0.34	0.40	0.35	0.43
Overall	0.47	0.40	0.47	0.40	0.46	0.39	**0.48**	**0.41**

表 1.精密工业（P）和召回（R）生的不同方法来自于好莱坞的新的数据集（动作之间没有时间关系）

Events	Hoai[6]		*ESM1*		*ESM-∞*		*SSM-∞*	
	P	R	P	R	P	R	P	R
AnswerPhone	0.64	0.32	0.65	0.36	0.64	0.32	0.64	0.43
HugPerson	0.46	0.29	0.49	0.33	0.48	0.33	0.51	0.33
Kiss	0.40	0.48	0.42	0.52	0.42	0.52	0.44	0.59
SitDown	0.36	0.36	0.38	0.38	0.39	0.38	0.39	0.38
Overall	0.47	0.36	0.48	0.40	0.48	0.39	**0.50**	**0.42**

表 2.精密工业（P）和召回（R）生的不同方法来自于好莱坞的新的数据集（动作之间没有时间关系）

个摄像机确定 k，通常会在 600-900 范围内。我们将在 6.4 节中给出一个更详细的关于 k 的分析。

对于我们评估的每个视频，我们首先应用我们的方法找到最优的分割和分类标识。在这一点上，每个段被分派给一个拥有开始和结束帧的特殊事件类。我们使用由[10]开发的一种二分的匹配方法，即根据事实环境（及参考环境注释）来校准检测结果。如果一个检测和一个真实事件匹配上了，则就被认为是真阳性，否则就是假阳性。这

6.2. 评价好莱坞的数据集

数据好莱坞是一个专注于现实人的行为的视频数据集。这些行为包括打电话，拥抱，接吻，坐下，下车，握手和站起等。这个数据集被分成两个不相交的子集，一个拥有 219 样本的训练集合和一个 211 样本量的测试集合。根据[6]，我们选择前四类作为被识别的动作，并将其它的看作空类。

由于好莱坞只包含预分割剪辑，为了我们的评估目的，我们通过连接原始数据及采摘的视频剪辑来创造了一个持续时间较长的视频剪辑。两个这样的数据集被创建了，都使用了来自用训练集和测试集的剪辑片段。对于第一个数据集，片段是被以一种随机串联的顺序选中，而第二个数据集，为了加强时间依赖性，一些片段被选中会来表现事件的相关性。具体来说，就是我们给数据中直插入一些具有一级（例如坐下，打电话）和二级（例如拥抱，接吻，嘴下）顺序依赖性动作。大约共有 40 多个视频样本以这样的方式形成，一半进入训练集一半进入测试集。

结果 我们通过标准精度和召回指标在好莱坞上报道了这个结果。如表 1 所示，当数据事件之间没有时间联系，所有的方法效果是一样的，而我们提出的方法（SSM-∞)稍微比别人的更好。然而，当时间依赖性被加到数据中的事件上时，所有的时间建模，方法优于基准，这表明事件信息对事件检测很有帮助。正如预期的，SSM-∞达到最好的结

果涉及精度和召回，展示了大量改进的基准线。在测试中 ESM-1 和 ESM-∞之间没有多大区别，因为该数据集被做作的呈现出中有简单的时间关系，并且电影剪辑中的场景具有显著的不同，为开采留下一些一点长的时间上下文。

6.3. 评价 SED 数据集

这个数据集是由放置在机场 5 个中不同位置的检测摄像机捕获而得到的。这数据集自 2009 年以来已被用在 TRECVID SED 评价轨道中，以支持在收集大量视频流数据中发展可视化视觉事件检测的新技术。它包含了人所从事自然活动的 10 件检测事件。在它们之中，7 个被使用在 TRECVID 2012 评估中，包括：CellToEar, Embrace, ObjectPut, Pointing, peopleMeet, PeopleSplitUp and PersonRuns。对于事件检测，这是

一个非常具有挑战性的数据集，由于许多混杂的事件，如高级别活动。相机视觉的改变，事件之间的各种差异偏差（如 PeopleMeet)和小的物质（如 CellToEar）（Fig. 5）。SED 的注释包括了时间范围和事件标签。

为了我们的评估我们使用 SED（约 100 小时的视频数据）的开发套件。这些数据被分成训练和测试相等的两部分。

结果 在 SED 上不同方法的结果被列于表 3。除精度和召回，检测成本率的分数（DCR）一个在 TRECVID 评价[16]被采用的性能指标，在表 3 中也提供了。根本的来说，DCR 是一个两个错误的线性组合：即遗漏检测和误报。它反映了这两种错误之间的一个折中通过在成绩中不用的称量他们。一个较低的 DCR 表现出更好的性能，关于这个指标的更多细节可在[16]中找到。

Events	#Ground truth	Chen[3]			Hoai[6]			*ESM*1			*ESM*-∞			*SSM*-∞		
		P	R	DCR	P	R	DCR	P	R	DCR	P	R	DCR	P	R	DCR
Cell2Ear	374	0.21	0.06	0.953	0.13	0.01	1.002	0.30	0.06	0.953	0.34	0.07	**0.933**	0.28	0.05	0.95
Embrace	479	0.13	0.27	0.835	0.12	0.24	0.856	0.14	0.27	0.833	0.15	0.28	0.812	0.17	0.34	**0.764**
ObjectPut	1898	0.33	0.05	0.985	0.43	0.01	1.001	0.39	0.05	0.969	0.44	0.06	0.941	0.45	0.09	**0.918**
PeopleMeet	1376	0.19	0.27	0.931	0.17	0.26	0.942	0.19	0.27	0.933	0.19	0.28	0.919	0.20	0.28	**0.913**
SplitUp	762	0.20	0.37	0.819	0.17	0.32	0.897	0.21	0.36	0.767	0.21	0.36	0.764	0.22	0.38	**0.715**
PersonRun	365	0.21	0.52	0.573	0.19	0.44	0.761	0.20	0.51	0.569	0.20	0.52	0.564	0.23	0.59	**0.499**
Pointing	2338	0.21	0.12	1.009	0.20	0.03	1.018	0.21	0.11	0.998	0.22	0.13	0.983	0.26	0.19	**0.958**
Overall	7592	0.19	0.15	N/A	0.20	0.17	N/A	0.20	0.19	N/A	0.24	0.22	N/A	**0.25**	**0.27**	N/A

表 3.精密工业（P），召回（R）和 SED 不同方法的 DCRS。需要注意的是一个较低的 DCR 分数表示一个更好的性能。整体 DCR 表现并不可作为由 TRECVID 提供的评估工具只能输出一个得分为每个单独的事件。

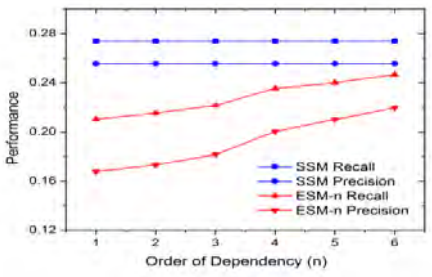

图 6.性能（精确度和召回）的比较通过使用不同长度的时间的上下文。

首先我们可以观察到，虽然 Hoai 的方法相比较于其它的方法在好莱坞中能够相当不错的执行，它却在 SED 上
不产生类似的结果。此外，相对于 Cell2Ear, ObjectPut 和 Pointing 事件具有非常低的召回，最困难的一点就是在数据集上进行检测。另一方面，在时间信息的帮助下，对于陈的方法 ESM-1 已
经达到类似的性能。通过进一步探索更长的时间背景上下文环境，ESM-∞ 和 SSM-∞优于陈的方法，清楚地表明了在数据序列中建模的效益比时间间的

events	*ESM*-∞			*SSM*-∞		
	P	R	DCR	P	R	DCR
Pointing	0.27	0.12	1.004	0.36	0.20	0.950
PersonRun	0.11	0.20	0.806	0.15	0.30	0.785
ObjectPut	0.39	0.06	0.944	0.45	0.09	0.933
Embrace	0.18	0.22	0.809	0.2	0.33	0.785

表 4. ESM-∞和 SSM-∞的基础上仅有部分性能地面实况（只在摄像机 SED-1）

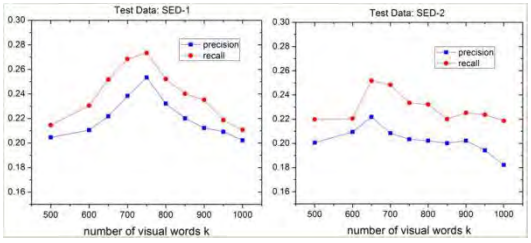

图 7. 我们的方法执行通过视觉单词上的两个亚群的数量变化（SED-1 和 SED-2）

相互作用更加复杂。SSM-∞在除了 Cell2Ear 的所有事件上产生了很好的效果。一些比较困难的事件例如 Pointing 和 ObjectPut 已经在 SSM-∞下基于基线得到显著的改善，通过我们的方法这表明了时间建模的功效。我们也注意到，我们的时间模型往往对那些其它表现出户没有明显的时间依赖关系的活动几乎没有什么效果，例如 Cell2Ear 和 PeopleRun。

6.4. 讨论

下面我们将在本文中为我们的方法提供更详细的分析来支持我们主要的权利要求。

远距离的时间依赖 为了证明在建模中利用远距离的时间依赖型的效果，我们比较我们提出的方法和正克的时空模型[22]方法的性能。图 6 清楚地说明了在数据中进行远距离时间环境建模的益处。随着时间依赖性范围的的增加，性能效果也不断地提高。

事实情况的影响 为了进一步了解在时间建模时地面实况产生的的影响，通过使用原始的地面真值集和一个降低的真值集来比较我们方法的实验结果这种方式，我们设计了了一个实验。具体的来说，我们从 SED 上的摄像机 1 重的事件注释中拿出两个事件（即 PeopleMeet 和 SplitUp），使用修改后的注释来运行 ESM-∞ 和 SSM-∞。请注意，这两个屏蔽掉的事件对 SEC 时间模型具有显著的贡献。从表 4 中，我们可以看到 ESM-∞如果没有地面实况明确表示的时间关系，就不能达到预期的效果。相比较而言，SSM-∞并没有因为不完美注视而受到严重的影响。这实验有利的支持我们的权利要求，在子事件粒度级别，对于 SM 而言，更有效地捕捉时间依赖性。

敏感性分析 k-均值聚类是用来生成我们的建模视觉序列。选择一个合适的 k 值能够帮助发现数据中的时间结构，K 不得不根据数据的复杂度进行选取。在更复杂的场景中，期待一个更大的 k 值。在我们的实验中，我们凭借经验确定每个相机上视觉词的数量。为了更好地了解 k 如何影响性能，我们真对于 k 评估性能的灵敏度。如图 7，当一些被精心调整的 k 期望获得更好的性能，在 SED 数据集上选择 k 在 600-800 之间就可以产生相当出色的表现。

7. 结论

在本文中，我们利用事件间的时间依赖性提出了一个联合分隔检测框架，其被认为是在视频中增强检测性能。使用 Squence Memoizer 来学习视觉词序列的依赖关系，它可以捕捉远距离的依赖关系和幂律特性。此外。我们模型的构造不依赖于事件的注释，并且还能够良好的处理空事件。我们已经在具有困难的数据集中出现了很有竞争力的效果，并证明了我们的方法由于国家的最先进的事件检测方法。

此外，请注意，我们只进行了有限的联合分隔（如真实事件，正确检测到事件之间的重叠）和在一个数据集（SED）中的主体中的识别误差分析（假阳性/阴性），当领域专家（NIST 和政府机关）制定个别误差的相对权重在 DCR 量度[16]形成时。

引用

[1] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. Computational linguistics, 18(4):467–479, 1992.

[2] M.-y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. 2009.

[3] Q.Chen, Y.Cai, L.Brown, A.Datta, Q. Fan, R.Feris, S.Yan, A. Hauptmann, and S. Pankanti. Spatio-temporal fisher vector coding for

surveillance event detection. In Proceedings of the21st ACMinternational conference on Multimedia, pages 589–592. ACM, 2013.

[4] K. Crammer and Y. Singer. On the algorithmic implementa-tionof multiclasskernel-based vector machines. The Journal of Machine Learning Research, 2:265–292, 2002.

[5] Q. Fan, R. Bobbitt, Y. Zhai, A. Yanagawa, S. Pankanti, and A. Hampapur. Recognition of repetitive sequential human activity. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 943–950. IEEE, 2009.

[6] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Confer-ence on, pages 3265–3272. IEEE, 2011.

[7] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In Computer Vision, 2003.Proceedings.NinthIEEEInterna tional Conference on, pages 1455–1462. IEEE, 2003.

[8] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In Computer Vision, 2009 IEEE 12th International Conference on, pages 1165–1172. IEEE, 2009.

[9] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 1, pages 166–173. IEEE, 2005.

[10] H. W. Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

[12] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In Computer Vision and Pattern Recogni-ion, 2007. CVPR'07.IEEEConference on, pages1–8. IEEE, 2007.

[13] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words.International Journal of Computer Vision, 79(3):299–318, 2008.

[14] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. Int. J. Comput. Vision, 77(1-3):103–124, May 2008.

[15] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2012 an overview of the goals, tasks, data, evaluation mechanisms and metrics. In Proceedings of TRECVID 2012. NIST, USA, 2012.

[16] P. Over, G. M. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, and G. Quénot.

Trecvid 2010–anoverview of the goals, tasks, data, evaluation mechanisms, and metrics. 2011.

[17] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In Computer Vision, 2009 IEEE 12th International Conference on, pages 1593–1600. IEEE, 2009.

[18] Q. Shi, L. Wang, L. Cheng, and A. Smola. Discriminative human action segmentation and recognition using semimarkov model. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

[19] K. Tang, F. Li, and K. Daphne. Learning latent temporal structure for complex event detection. In Computer Vision and Pattern Recognition (CVPR),2012 IEEE Conference on, pages 1250–1257. IEEE, 2012.

[20] D. Tran and A. Sorokin. Human activity recognition with metric learning. In Proceedings of the 10th European Conference on Computer Vision: Part I, ECCV '08, pages 548–561, Berlin, Heidelberg, 2008. Springer-Verlag.

[21] F. Wood, C. Archambeau, J. Gasthaus, L. James, and Y. W.Teh. A stochastic memoizer for sequence data. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1129–1136. ACM, 2009.

[22] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W.Teh. The sequence memoizer. Communications of the ACM,54(2):91–98, 2011.

[23] Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In Computer Vision–ECCV 2012, pages 707–721. Springer, 2012.

[24] G. K. Zipf. Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press, 1932.