指导教师：杨涛　　提交时间：2015/3/29

The task of Digital Image Processing

数字图像处理

School of Computer Science

No: 1

姓名：张雨超

学号：2012302493

班号：10011205

基于推断时间实例标签的视频事件检测

Kuan-Ting Lai[§†], Felix X. Yu[‡], Ming-Syan Chen[§†] and Shih-Fu Chang[‡]
[§]Graduate Institute of Electrical Engineering, National Taiwan University, Taiwan
[†]Research Center for IT Innovation, Academia Sinica, Taiwan
[‡]Department of Electrical Engineering, Columbia University, USA
{ktlai, mschen}@arbor.ee.ntu.edu.tw, {yuxinnan, sfchang}@ee.columbia.edu

摘要

视频事件监测允许基于事件对视频内容进行智能检索。传统方法是从视频帧或者镜头中提取特征，通过量化并合并特征，使整个视频形成一个单一的向量表示。这虽然是一种简单有效的方法，但是最后合并的步骤可能会导致局部时间信息的丢失，而这种信息对于在长视频中指明哪一部分能表明事件的出现是很重要的。在这项工作中，我们提出了一个新颖的基于实例的视频事件检测方法。我们把每个视频都描绘成许许多多的实例，分别定义为不同时段的视频段。目的在于通过基于视频级的标签来进行实例级的事件检测模型的学习。为了解决这个问题，我们提出了一个大胆的构想，就是把实例标签当做隐含的潜变量，同时推断实例标签以及实例级分类模型。我们的框架推断出最佳的解决方案，即假设正视频有大量的正实例而负视频则拥有最少的。此外，大型视频事件数据集上的大量实验展示出了显著地性能提升。需要指出的是，该方法在解释检测结果上也很有用，方法是定位视频中和正检测有关的时间段。

图 1 对提出框架的解释。生日宴会事件可以通过包含“生日蛋糕”和“吹蜡烛”的实例 识别出来。我们的方法只基于视频级标签便可以同时推断出隐含实例标签以及实例级的分类模型（分离超平面）。

1. 介绍

视频事件检测在许多应用中都是很有用的，比如视频搜索、消费类视频分析、广告个人定制以及视频监控，仅列举几例[15]。已经有许多视频事件检测方法被提出，包括巨大边界方法，图形模型以及基于知识的技术等等[9]。最常用的方法是把视频作为一个全局词袋(BoW)向量[18]。BoW法可以被分为三步：首先，从视频帧或者视频段中提取局部特征（视觉，语音或者属性）。接着，这些特征由经过学习的代码本或者字典量化[24]。最终，合并这些量化特征，以形成一个全局的向量表示。

元素代表一个视频作为一个单一的载体是简单而有效。不幸的是，很多信息可能会丢失在最后汇集一步，从而导致不令人满意的性能。事实上，一段视频是由多个"实例"，如框架和镜头组成。某些情况下包含事件正在考虑的关键的证据。例如，像"生日派对"事件可能由帧包含蛋糕和蜡烛而较好地检测，"跑酷"则可以由人在大街上跳上跳下而很好地检测。

最近的研究中所示，人类可以很好利用关键证据并只由较短的视频片段认识时间[2]。将视频特征提取成单一的聚合形式可能不能在大量实例之上充分利用如此丰富的线索。直观地说，通过考虑视频的实例，更有特色的事件模式是可以学会的并因此可以达到更好的事件识别。

出于上述事实，我们研究了基于实例的视频分类，如图一所示。每个视频包含多个的"实例"，定义为不同时间长度的视频片段。实例的定义可能是灵活的——他们可以是视频帧，固定长度的视频，基于内容变化的视频片段或者整段视频。在本文中，我们建议考虑多粒度，即不同视频长度的实例。这给了我们在建模视频事件的不同时间尺度上的灵活性： 一些可能是短时间的而别人可能需要很长的时间间隔。图2 在一段视频中显示了实例多粒度。我们的目标是学会一种实例级别的事件检测模型，同时假设由于标注实例的高昂成本，实例标签是不可得的。 基于对多实例学习 (MIL) 的常规方法可能似乎是很自然的选择。但由于它的结束简化模型在将实例预测转移到袋，它不是一个令人满意的解决方案，我们将会在稍后展示并且评估它。

为了解决这个具有挑战性的问题，我们建议大边缘框架。它把该实例标签作为隐式t变量，同时推断隐含实例标签以及实例级分类模型。我们的主要假设是正的视频通常具有一大部分正的实例，而负的视频几乎没有。提出的方法不仅导致事件检测结果更准确，也学会了实例级探测器，解释什么时候和为什么某一事件发生在视频中。

我们的论文包括以下几个重要部分：

- 我们提出了一种新颖的基于实例的视频事件检测方法 （第 3 节）。
- 基于大缘学习框架，我们开发一种算法，可以仅从视频级别标签同时推断实例标签和实例级别的事件检测模型（第 3 和第 4 节）。
- 大量的实验评估说明了在大规模视频数据集上我们方法的超群性能。

2. 相关工作

2.1. 视频事件检测

视频事件检测是计算机视觉中的广泛研究的课题。文献[9]中对技术发展水平进行了一个很好的调查。一般来说，视频事件检测系统可以分为三个阶段： 特征提取、 特征量化和汇集，训练和识别。

以前的研究的一个重点是设计新特征。包括低层特征的视觉特征 [6，12]、 行动特点 [22]、 音频功能 [14] 和中级首季包括概念特征属性 [20] 等。我们在改进事件识别模型上也花了很大努力，比如在[9]中介绍的基于大边缘的方法，图形模型和一些基于知识的技术。然而，大多数之前的方法依赖于一个全局的向量来表示一个视频。全局的做法忽略了事件的重要的本地信息。最近一些研究人员试图解决这一问题，并提出几种新的算法。唐等人[19]将视频片段作为潜变量，并通过持续变量隐马尔可夫模型来表示的事件。曹等人[3]提出场景对齐汇集，它把视频拍摄成不同的场景，并且在每个场景中汇集局部特征。李等人[11]提出了动态汇集，它采用各种策略，以视频分为基于时间段的结构。不同于他们集中于开发汇集的时间架构方法，我们的框架的重点是学习"实例"标签。所提出的方法还可以看成对上述汇集策略的补充，即视频实例可以通过动态汇集或场景对齐汇集来表示。

2.2. 多实例学习

为了在一段视频中使用局部模式，一个容易利用的学习方法是多实例学习 (MIL)[7]. 在多实例学习中，训练数据按包提供，并且只提供包级的标签。一个包被标为正的当且仅当包中的一个或多个实例是正的。在计算机视觉领域，MIL已在场景分类应用[13]，基于内容的图像检索[27]，和图像分类[5]。两个最流行的算法MIL是MI-SVM和MI-SVM[1]。第一个算法强调搜索最大边界的超平面来分离正负实例，而第二个算法在最优化迭代中从每个正包中选出最有代表性的正实例，并且把注意力放在包分类上。

在事件检测中，视频可以被看作包含多个实例的包。由于标签仅在视频级提供，MIL算法可直接应用。然而，MIL的现有算法不适合于视频事件分类。一个限制是MIL依赖于预测的单一实例（通常在最大函数上计算），使得方法误报异常值非常敏感。另一个缺点是，它假设负包中没有正实例，导致对于复杂事件会

图2.图示具有不同时间长度的多个粒状实例。在我们的框架中，最小粒度的实例是一帧，并且最大粒度的实例是整个视频。每个实例都由包中包含的帧级弓表示。

产生不稳定结果。

2.3. 标签比例学习

几种方法已经被提出来解决的MIL的局限性。陈等人[4]提出通过实例相似测算把包嵌入实例空间。张等人[26]提出了既考虑局部（实例）又考虑全局（包）特征向量的新方法。MIL的另一个产物是标签比例学习。在标签比例学习中，学习者可以访问每个包正实例的比例。和MIL相比，LLP 能够产生更稳定的结果，因为模型并不只是依赖于每个包中的单一实例。LLP的许多方法已经被研究出来了 [16,25]。

我们的视频识别算法受到了SVM的启发Ap-SVM or ∝SVM) [25], 他通过大边缘框架中已知的标签比例详尽地为潜在的未知实例建模。∝SVM被证明优于其他替代品。和∝SVM不同A 在视频分类中A 准确的标签比例是未知的。我们的主要假设是，实例中的正视频的大部分应该是正的，而在负的视频少数情况可能是正的。我们还考虑多时间粒度，以形成实例，从而导致在视频分类显著性能改进。

3. 提出的方法

设置。　假设我们拥有训练用的数据集$\{V_m\}_{m=1}^{M}$ 考虑一个单独的事件，即在每一个Vm 中有Nm个实例$\{\mathbf{x}_i^m, y_i^m\}_{i=1}^{N_m}$, 其中 $\mathbf{x}_i^m$ 是第m个视频第i个实例的特征向量, 同时相应的事件标签是 $y_i^m \in \{1,1\}$。在这里，如果事件实例为正，那么$y_i^m = 1$，否则为-1。正如第一节指出的，对于大多数情况，实例标签是未知的，而监督信息只是在视频层上被提供。因此，我们建议学习仅基于视频级监督信息的一个实例级分类模型。

在3.1节我们提出在简单实例下用公式近似∝SVM ，其中，每个视频正实例的比例是已知的。在3.2节，我们我们说明如何该方法延伸到真实世界的情况下，在其中只有视频二进制级别事件标签给出。在3.3节，我们建议使用改进的视频分类多粒度的实例。

3.1. 通过实例比例进行视频检测

在这一节，我们考虑一个简单的例子，每一个视频中，我们认为正实例的比例 $\{P_m\}_{m=1}^{M}$是已知的。在这里，$P_m \in [0,1]$ 是第m个视频Vm的正实例。目标是培养一个实例级的分类器，以个别的实例进行分类。我们提出学习大边缘的事件分类模型(w, b)，也就是说，当$(\mathbf{w}^\top \mathbf{x} + b) > 0$,是预测x为正，而当 $(\mathbf{w}^\top \mathbf{x} + b) \leq 0$时预测x为负。为了解决这个问题，我们建议联合推断实例标签和预测模型。我们的公式和∝SVM　[25]相一致。它试图去寻找一个和所给标签比例兼容的大边界分类器。我们注意到，如果给出实例标签 $\mathbf{y}^m$ ，第m个视频的正实例比例$p_m(\mathbf{y}^m)$ 可以表达成:

$$p_m(\mathbf{y}^m) = \frac{\sum_{i=1}^{N_m} I_{(y_i^m=1)}}{N_m}, \tag{1}$$

其中 $\mathbf{y}^m = [{y_1}^m, \cdots, {y^m}_{N_m}]$. $I_{(\cdots)}$ 表示指示函数，当参数为真值时用1表示，而当参数为假值时用0表示。上式和如下表达式是等价的:

$$p_m(\mathbf{y}^m) = \frac{\sum_{i=1}^{N_m} y_i^m}{2N_m} + \frac{1}{2}, \tag{2}$$

分类模型的参数和未知的实例标签可以通过最优化以下的目标函数而共同学习:

$$\min_{\{\mathbf{y}^m\}_{m=1}^{M}, \mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{m=1}^{M}\sum_{i=1}^{N_m} L\big(y_i^m, (\mathbf{w}^\top \mathbf{x}_i^m + b)\big)$$
$$s.t. \quad p_m(\mathbf{y}^m) = P_m, \quad m = 1, \cdots, M. \tag{3}$$

第一项是经典的SVM项，目的是找到一个两个类的最大边界分离超平面 。第二项$L(\cdot)$ 是实例标签和预测的经验损失函数。

我们所提出的框架允许为$L(\cdot)$选择不同的损失函数。在本文中，我们L(·)使用铰链损失函数，即：

$$L\big(y_i^m, \mathbf{w}^\top \mathbf{x}_i^m + b\big) = \max(0, 1 - y_i^m(\mathbf{w}^\top \mathbf{x}_i^m + b)). \quad (4)$$

综上所述，该框架试图找到一个大比分的分类，与给定的标签的比例不兼容。作为特殊情况，如果我们已知所有的实例标签$y_i{}^m, m = 1, \cdots, M, i = i, \cdots, N_m$，框架就变成了经典的监督SVM。

3.2. 处理未知比例

在前面的章节我们讨论了当实例标签 $\{P_m\}_m{}^M{}_{=1}$ 是已知的情况。但是，在视频分类中，我们只知道视频级的二进制标签 $\{Y_m\}_m{}^M{}_{=1}$, 其中 $Y_m \in \{-1, 1\}$ 是V_m的视频级标签。

为了解决这个问题，我们的主要假设是，各积极视频包含多的正实例，而每一个负视频包含很少或没有正的实例。具体地讲，我们提出以下修改的公式以实现上述的假设：

$$\min_{\{\mathbf{y}^m\}_{m=1}^M, \mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{m=1}^{M} \sum_{i=1}^{N_m} L\big(y_i^m, (\mathbf{w}^\top \mathbf{x}_i^m + b)\big) + C_p \sum_{m=1}^{M} |p_m(\mathbf{y}^m) - P_m| \quad (5)$$

$$s.t. \quad P_m = \begin{cases} 1 & \text{if } Y_m = 1 \\ 0 & \text{if } Y_m = -1 \end{cases}, m = 1, \cdots, M.$$

第一个修改时把等式3的严格限制转移到目标函数：第三项是一个损失函数用来补偿目标正实例比例P_m和估计的比例$p_m(\mathbf{y}^m)$。第二，我们把正视频的正实例比例设为1，负视频则设为0。在这种设置中，框架鼓励正视频中大比例的正实例，同时抑制负视频中的正实例。C_p是控制我们假设强度的参数，实际上，C_p可以根据交叉验证进行调整。

3.3. 实例多粒度

一个还没有解答的关键问题是怎么为每个视频设计实例实例可以是帧，或者镜头，乃至视频片段甚至整个视频。不同时长的实例对于识别不同的时间是有效的。比如说，生日聚会可以仅仅通过包括蛋糕和蜡烛的帧来识别。而和运动有关的动作，比如技巧滑板和跑酷，则更容易通过视频片段来检测。

通过观察动机，我们认为根据不同长度的时间间隔多粒度的实例。多颗粒实例的特征表示是通过集中的局部特征与具体的时间长度段级BoW获得。需要注意的是视频BoW是在我们的框架中的一个特例。

最初的∝SVM 同等的处理所有实例并且不能够区分多粒度实例。因此，我们开发了一个新的公式，可以分配权重不同的颗粒的实例。我们提出的公式如下介绍。

假设我们拥有K粒度。第m个视频中，第k级粒度表示为 N_k^m，接着，我们可以定义一个如下的标签向量 $\mathbf{y}_k{}^m = [(y_1)_k{}^m, \cdots, (y_{N^m})_k{}^m]$。第m个视频中第k级粒度的第i个标签为$(y_i)_k{}^m$。第k个粒度的权重定义为 t_k。

因此，我们写新的比例函数 $p_m(\mathbf{y}_1^m \cdots \mathbf{y}_K^m)$ 。

$$p_m(\mathbf{y}_1^m \cdots \mathbf{y}_K^m) = \frac{\sum_{k=1}^K t_k(\mathbf{1}^\top \mathbf{y}_k^m)}{2\sum_{k=1}^K t_k N_k^m} + \frac{1}{2}. \quad (6)$$

第m个视频的总实例数N_m现在是实例的所有粒度的加权总和 $\sum_{k=1}^K t_k N_k^m$，在本文中，实例的来自第k个粒度的特征通过包含帧对的平均的BoW表示来计算，而权重仅仅设置为所包含的帧数。我们把第m个视频的第k个粒度的第i个特征向量记做$(\mathbf{x}_i)_k^m$，则权重版本的等式5如下所示：

$$\min_{\{\mathbf{y}^m\}_{m=1}^M, \mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C_p \sum_{m=1}^{M} |p_m(\mathbf{y}_1^m \cdots \mathbf{y}_K^m) - P_m| + C \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{N_k^m} t_k L\big((y_i)_k^m, (\mathbf{w}^\top (\mathbf{x}_i)_k^m + b)\big)$$

$$s.t. \quad P_m = \begin{cases} 1 & \text{if } Y_m = 1 \\ 0 & \text{if } Y_m = -1 \end{cases}, m = 1, \cdots, M. \quad (7)$$

求解上述方程是一个具有挑战性的问题，因为它是一个不能在多项式时间内解决NP-难组合优化问题。在下一节中，我们将解释我们的战略和阐述优化过程中的每一步。

4. 过程优化

为了解决等式7的问题, 我们应用可选择的优化找到一个局部的次优解：

- 首先我们安装实例标签$\{\mathbf{y}^m\}_{m=1}^M$，并且解决w和b。通过安装 $\{\mathbf{y}^m\}_{m=1}^M$，优化问题变成了一个经典的权重SVM。

$$\min_{\mathbf{w}, b} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{N_k^m} t_k L\big((y_i)_k^m, (\mathbf{w}^\top (\mathbf{x}_i)_k^m + b)\big).$$

- 第二我们安装$\mathbf{w}$和b并且更新实例标签$\{\mathbf{y}^m\}_{m=1}^M$
  接着问题变成了:

$$\min_{\{\mathbf{y}_m\}_{m=1}^M} \quad C\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{N_k^m} t_k L\big((y_i)_k^m, (\mathbf{w}^\top(\mathbf{x}_i)_k^m + b)\big) + \; C_p\sum_{m=1}^{M}|p_m(\mathbf{y}_1^m\cdots\mathbf{y}_K^m) - P_m|. \tag{8}$$

由于每个视频V_m对目标有独立贡献，我们可以一次一个视频的优化等式(8)。视频V_m 的优化过程如下。

我们把实例标签$(y_i)_k{}^m$在每个视频V_m中的值都设为-1，并且计算每一个经验损失增长$(\delta_i)_k{}^m$，通过转置每一个实例标签 $(y_i)_k{}^m$ 的值为1。$(\delta_i)_k{}^m$ 的值可以如下计算:

$$(\delta_i)_k^m = \big(1-(\mathbf{w}^\top g_k(\mathbf{x}_i^m)+b)\big)_+ - \big(1+(\mathbf{w}^\top g_k(\mathbf{x}_i^m)+b)\big)_+,$$

其中函数$(x)_+ = \max(x, 0)$。

一旦所有经验损失增长$(\delta_i)_k{}^m$计算好，对所有的权重损失值$t_k(\delta_i)_k{}^m$按照降序进行排列。这部分和原始的$\propto$SVM算法是一样的。但是，在我们的公式中寻找最小总损失是一个不同的问题。在$\propto$SVM算法中实例标签被一个个的转置来计算比例损失增长，并且拥有最小总损失的实例的数量也被选择为翻转。当实例有不同权重的时候，有不止一种组合可以得到某种比例。在我们的框架中，我们采用一种可以指数到线性时间复杂度范围内求解的贪心算法来搜索一个次优解。

我们提出的优化过程在算法1中展示。目标函数在我们的优化过程重视非增长的。当目标函数减小小于一定的阈值的算法停止，在我们的实验中被设置为10^{-2}据经验，优化过程收敛速度在短短的几十次迭代内就可以完成。虽然上述的方法是基于线性的大裕度的框架，它可以很容易地扩展通过施加内核技巧解决配有固定实例标签的内核方案（W，B）。

Algorithm 1 Optimization Procedure

Input: $k = 1\cdots K, m = 1\cdots M$
video label $Y_m \in \{1, -1\}$.
instance $\mathbf{x}_k^m$, instance weight $t_k \in \mathbb{R}$.
proportion $P_m = 1$ if $Y_m = 1$, $P_m = 0$ if $Y_m = -1$.
convergence threshold $\theta = 0.01$.
Initialization:
$(y_i)_k^m \leftarrow Y_m, i = 1\cdots N_k^m, k = 1\cdots K, m = 1\cdots M.$
repeat
  fix $\mathbf{y}_k^m$ and solve $\mathbf{w}$ and b.
  set cost reduction $C_R \leftarrow 0$.
  for $m = 1\cdots M$ **do**
    $(y_i)_k^m \leftarrow -1, k = 1\cdots K, i = 1\cdots N_k^m$
    compute all $(\delta_i)_k^m$ for $(y_i)_k^m$ (Eq. 9)
    sort $(y_i)_k^m$ by $t_k(\delta_i)_k^m$ in descending order
    for sorted $(y_i)_k^m$ **do**
      flip $(y_i)_k^m$, calculate the loss reduction incrementally.
    end for
    select maximum loss reduction.
    flip the labels to get max loss reduction, and update C_R.
  end for
until convergence ($C_R < \theta$)
Output: $\mathbf{w}, b, \mathbf{y}$

5. 讨论

5.1. 视频级事件检测

在上一节中，我们提出了基于视频级标签学习的事件检测模型。我们的方法的一个固有的优点是，它可以自然地发现它们支持的特定事件的存在的主要依据。排名靠前的由我们的方法所选择的16个证据在图4和图8中展示。某些选定的单帧实例是强烈的证据，其中人类可以通过观看这些帧确认目标事件的存在。

为了在视频级上进行事件检测，我们可以在测试视频的所有实例上使用实例分类器。视频级检测得分接着可以通过对所有实例得分进行加权平均而获得。直观地说，含有较多的正实例的视频往往为正的概率更高。我们将在后面通过实验展示我们的方法可能会导致视频级事件检测显著的性能提升。

5.2. 计算成本

由于我们使用了一个类$\propto$SVM算法，计算代价（线性SVM解法） 为$\mathcal{O}(N\ \log\ \max_m(N_m))$,其中 $\max_m(N_m)$是m个视频的最大实例,并且N实例总数。表达式可以写作$\mathcal{O}(\mathcal{V}\mathcal{T}\ \log\ \max_m(N_m))$，其中$\mathcal{V}$是视频总数并且$\mathcal{T}$是每个视频的平均实例数。由于$\mathcal{T}\ \log\ \max_m(N_m)$ 可以被看作是一个常数，计算复杂度和线性SVM算法下基于视频的事件分类器一样。

在实践中，多种技术可应用于提高的计算时间。 比如说，本框架可以被大幅改进，通过在内循环内求解SVM算法。一种途径是利用有Shilton等人提出的热启动和部分活跃集的方法[17]。另一种方法是运用使用详尽特征映射的非线性内核，以使我们方法的复杂度即使在非线性内核的情况下也能成为线性。

5.3. 异构实例学习

在上一节中，我们考虑用相同底层特征表示的实例多粒度。在实践中，情况可能会用不同的表述。例如我们可能会通过图像/音频/动作分别代表实例的特征。在这种情况下，我们提出的方法可以在不做什么修改的情况下应用来为每种特征表示学习一个分类模型。我们也可以共同学习经过修改过的目标函数的分类模型。我们把这个课题留给我们将来的工作。

6. 实验

数据集。 为了评估我们的框架中，我们进行了实验三个大型视频数据集：TRECVID的多媒体事件检测（MED）2011年，MED2012 [15]和哥伦比亚消费者视频（CCV）[10]的数据集。我们所有的实验都是基于SIFT特征和线性SVM。

特征。 在本文中，我们选择了SIFT[12]作为底层局部特征进行初步评估。请注意，通过使用融合技术我们的方法可以很容易地扩展到包括多种功能。例如，我们可以训练不同的基于实例的检测模型，各自独立拥有和熔断器检测探测器的成绩使用不同的功能，最终事件检测。此外，通过采用多种功能，我们可以为每个视频事件发现独特的线索，如动作，颜色，音频。

设置。 对于每个视频，我们每2秒提取一帧。 每一帧的尺寸为320 × 240，并且SIFT特征由VLFeat库以10像素为步长密集提取。帧特征被量化为5000词袋向量。帧级SIFT词袋作为实例特征向量。线性SVM工具[8]被用于当实例标签y被安装时求解$\mathbf{w}$和b。基于交叉验证的成本参数C和C_p从规模$\{0.01, 0.1, 1, 10, 100\}$中选择。

基准线。 我们评估的数据集中在四个基本算法：MI-SVM，MI-SVM[1]，视频词袋以及单帧实例比例SVM（αSVM或P-SVM）[25]。在MI-SVM和MI-SVM都使用多个实例学习。正如在2.2节介绍的那样,MI-SVM集中在实例级，而MI-SVM侧重于袋级分类。在实践中，MI-SVM推断实例标签迭代，同时迫使负视频的所有实例为负。MI-SVM选择与每个迭代中得分最高的每一个正视频的一个正实例，并迫使负视频的所有实例为负。我们采用MILL库[23]并且改性它以使其获得更好的性能。对于∝SVM，我们就像在3.2节中一样，设置一个未知比例 ，并且只选择单帧作为实例。

[1] https://github.com/felixyu/pSVM

ID	MED 2011 Events	ID	MED 2012 Events
1	Attempting board trick	16	Attempting bike trick
2	Feeding animals	17	Cleaning appliance
3	Landing a fish	18	Dog show
4	Wedding ceremony	19	Give directions to location
5	Woodworking project	20	Marriage proposal
6	Birthday party	21	Renovating a home
7	Changing a tire	22	Rock climbing
8	Flash mob gathering	23	Town hall meeting
9	Getting vehicle unstuck	24	Win race without a vehicle
10	Grooming animal	25	Work on metal craft project
11	Making sandwich		
12	Parade		
13	Parkour		
14	Repairing appliance		
15	Sewing project		

Table 1. 在TRECVID MED 2011和2012中定义的25个事件

6.1. 哥伦比亚消费者视频(CCV)

哥伦比亚消费者视频(CCV)基准定义20个事件，并包含从YouTube上下载的9317个视频。事件名称和测试拆分可以在原始论文[10]中找到。

在选择多粒度情况而言，我们评估CCV数据集在不同情况下的组合。凭经验，我们发现，更粒度导致更好的性能。 例如，使用4粒度（单个帧，3帧拍摄，和5-帧拍摄和整个视频）的结果实现mAP0.436的最佳结果。然而，如果增加粒度的数目将会导致更高的计算成本。考虑到时间和性能之间的权衡，在本文所有的实验中，我们只使用两个粒度：单幅和整个视频实例。

该实验的结果示于图3。MI-SVM和MI-SVM不如标准的视频级词袋法。这是由于MIL的限制性假设，它的重点是在每个视频搜索一个最有代表性的实例，并且把所有在负视频中的实例当做是负实例。相反的是，∝SVM没有做这样的假设并且优于视频词袋。我们的方法通过考虑多粒度实例进一步提高了性能，分别相对于 视频词袋法和∝SVM法提高了10.2%和4.8%。

6.2. 关于TRECVID MED12

MED12数据集包括25个复杂事件和5816个视频。MED 11和MED 12中事件名称被列在了表1。我们把三分之二的数据作为训练集（3878个视频），并使用剩下的作为测试集（1938个视频）。抽取的帧中的每个视频的平均数量是79.4，并且一个事件上的单一的Intel Xeon的CPU@2.53GHz的平均学习时间为约40分钟。实验的结果示于图5。结果的比较中发现类似的观察到的C-CV数据集。MI-SVM和MI-SVM不如标准的视频级词袋法。

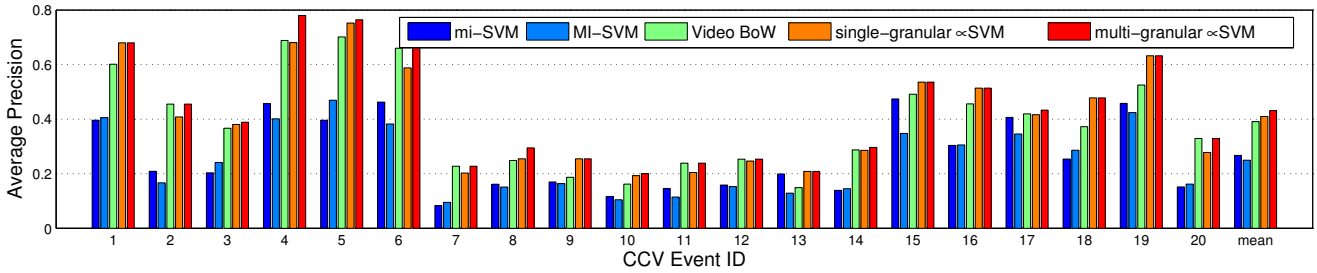


图3。 在哥伦比亚大学消费者视频（CCV）的20个复杂事件的实验结果。平均接入点是0.26（MI-SVM），0.25（MI-SVM），0.39（视频词袋法），0.41（单粒αSVM）和0.43（多颗粒αSVM）。

(a) 自行车技 (b) 狗狗秀 (c) 岩石攀登 (d)市政厅会议 (e) 田径比赛

图4。 在MED12中选择的前16个正视频帧。 该方法可以为每个事件成功检测重要的视觉线索。例如，最高级实例“赛跑”是关于田径场的。

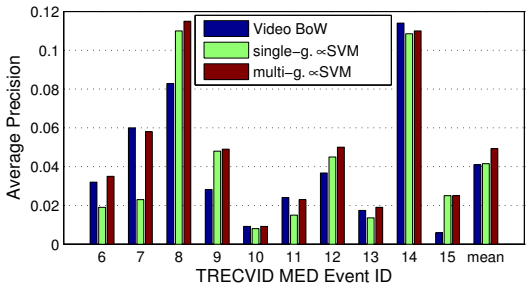


图6。MED 2011事件6到15的得分

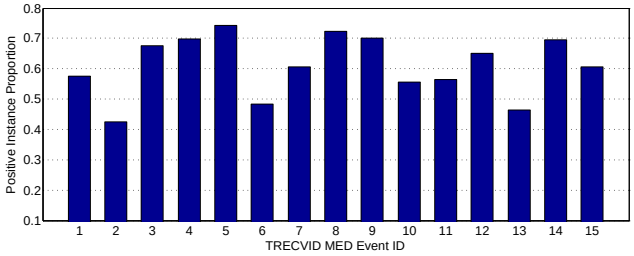


图7。在MED11活动中通过我们的方法进行学习得到的正实例比例。

我们的方法各自相对视频词袋法和αSVM算法性能提高了21.4%和9.68%。正如前面提到的，我们的方法对精确定位特定事件的局部片段有着很大的贡献。图4展示了在视频中自动选择检测为正的关键帧，这对于解释检测结果是很有用的。

6.3. 关于TRECVID MED11

在这个实验中，我们按照TRECVID MED大赛的官方数据分割。NIST提供MED11三个数据分割：事件收集（EC），发展集（DEVT）和测试集（DEVO）。T他的事件集合包含了超过15事件2680训练视频。设置10,403视频的DEVT为参赛者以评估他们的系统而发布。最终的性能在拥有32061测试视频DEVO的上进行评价。每个视频取出的帧的平均数目是59.8，并且一个事件上的单一的Intel Xeon的CPU@2.53GHz的平均学习时间约6小时

实验结果在图6中展示。从事件1至事件5，仅事件6至事件15展示出来。该∝SVM算法优于视频词袋的“快闪族聚会”，“获得车辆脱胶”和“游行”，但产生了其它事件更坏的结果。这是一个有趣的发现，因为它证实，都需要代表不同的事件不同长度的实例。我们的方法在该实验中优于其他方法20%左右。图7说明为每个事件学习的正实例的比例。最优正实例的比例可以低至42.6%，尽管在目标方程设置为100%。使用我们的方法进行学习的一些高级帧实例在图8展示。

7. 结论

我们提出了一个新的方法，通过同时推断例如标签和学习实例级事件检测模型来进行视频事件检测。我

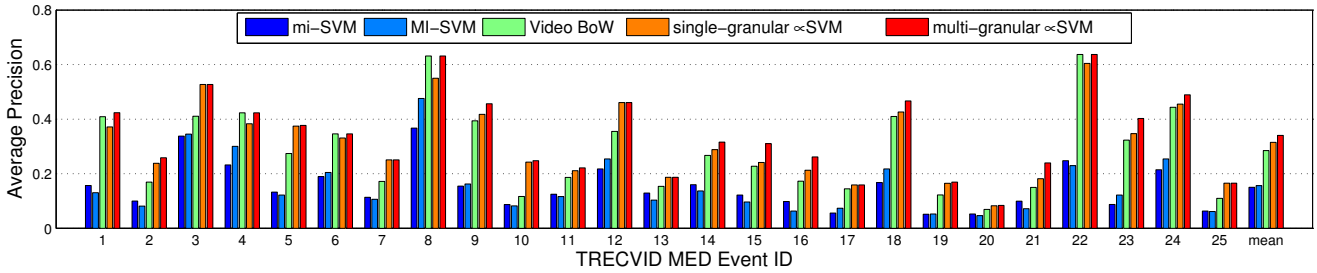
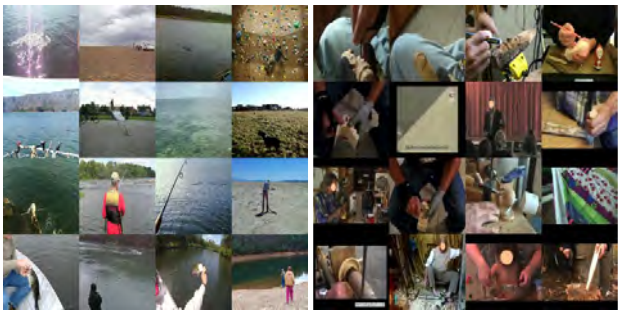


图 5。在TRECVID MED12个视频25集复杂事件的评估结果。平均接入点是0.15（MI-SVM），0.16（MI-SVM），0.28（视频词袋法），0.31（αSVM）和0.34（我们的方法）。



(a) 钓鱼　　(b) 木工项目

图8。通过我们的算法，从TRECVID MED11的一些事件中选择出最好的16个关键正视频帧。

们提出的算法考虑的实例多粒度，充分利用局部和全局的模式，以达到最佳效果，因为清楚地表明了广泛的实验。所提出的方法也可以通过局部化具体的表示该事件出现的时间帧/段提供检测结果的直观解释。

参考文献

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[2] S. Bhattacharya, F. X. Yu, and S.-F. Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *ICMR*, 2014.

[3] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, and J. R. Smith. Scene aligned pooling for complex video recognition. In *ECCV*. 2012.

[4] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *PAMI*, 28(12):1931–1947, 2006.

[5] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *JMLR*, 5:913–939, 2004.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[9] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *IJMIR*, pages 1–29, 2012.

[10] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011.

[11] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[13] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, 1998.

[14] P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *PAMI*, 116:374–388, 1976.

[15] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2013*. NIST, 2013.

[16] S. Rüeping. SVM classifier estimation from group probabilities. In *ICML*, 2010.

[17] A. Shilton, M. Palaniswami, D. Ralph, and A. C. Tsoi. Incremental training of support vector machines. *IEEE Transactions on Neural Networks*, 16(1):114–131, 2005.

[18] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

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

[20] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*. 2010.

[21] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 34(3):480–492, 2012.

[22] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[23] J. Yang. Mill: A multiple instance learning library, 2009.

[24] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[25] F. X. Yu, D. Liu, S. K., T. Jebara, and S.-F. Chang. ∝SVM for learning with label proportions. In *ICML*, 2013.

[26] D. Zhang, J. He, L. Si, and R. D. Lawrence. Mileage: Multiple instance learning with global embedding. In *ICML*, 2013.

[27] Q. Zhang, S. A. Goldman, W. Yu, and J. E. Fritts. Content-based image retrieval using multiple-instance learning. In *ICML*, 2002.