视觉的节奏和节拍

ABE DAVIS, Stanford University

MANEESH AGRAWALA, Stanford University

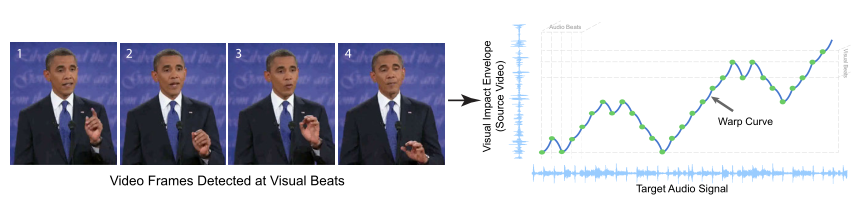
1

Figure 1. 我们搜集无节奏意识的视频片段,分析其中的节奏. 然后, 我们按照某一音乐中精确对齐的节拍, 把这些片段规整(warp)成舞蹈表演. 在图片左边, 是一段视频(2012 presidential debate video [WSJDigitalNetwork 2012])中检测出来的在拍点上的4个连贯的视频帧. 这些视觉拍子对应重复上上下下的手势动作. 图片右边的曲线表示了折叠伸缩视频的过程, 可以综合出一个由原视频中随机片段组成得到的鬼畜(dance)视频

摘要:

在分析视频的动作中, 我们得到一种类似于音乐节奏的视觉节奏. 并且展示了视觉节奏与音乐对齐在一段舞蹈中的效果. 我们工作的核心是视觉节拍——可以在时间刻度上调整,从而控制视觉节奏的动作模式. 我们可以通过将视觉上的拍子规整成与音乐对应, 创造或操纵一段视频. 通过这种方式, 我们为＇控制音视频节奏的同步性＇这一技术做了若干应用：我们可以改变舞蹈的配乐，可以把不规则的动作规整成一段与音乐对应的舞蹈，或者可以搜集一些凑巧像是舞蹈的动作片段，合成一段表演．

1引言

音乐和舞蹈通过节奏的概念密切相关，它描述了一些事件——例如乐器的声音或身体的运动——在时间刻度上是怎样分部的．节奏在某些方面是一个非常直观的概念：婴儿早在六个月大就可以识别并跟随的基本的节奏了[Cirelli et al . 2016; Repp and Su 2013]，甚至一些动物——比如某些鹦鹉和大象，会随着简单的音乐有规律的动， [Patel and Demorest 2013; Patel et al . [n. d.]]．然而，对于节奏的量化的任务从来都不简单，一直是音乐[Böck and GerhardWidmer 2013; Dixon 2006; Ellis 2007; Goto 2002; Grosche et al . 2010; Huetal . 2017;Lerch2012]和舞蹈[BrickandBoker2011;Dyaberiet al . 2006]研究领域中的主题．我们的工作建立在那些，从视频中发现对应于音乐节奏的，我们称之为视觉节奏（visual rhythm），的研究之上．相对于音乐节奏捕捉短时间的声音排列，视觉节奏捕捉短时间的动作的排列．我们关注研究那些包含着能被标记为与舞蹈相关的结构的动作．

我们的核心假设是，音乐和舞蹈的特点是在听觉信号或视觉信号上有足够的节奏结构（rhythmic structure）．我们对这种结构的研究是基于视觉节拍的概念——与音乐节拍对齐的，组成舞蹈片段的一系列视觉事件．视觉和听觉节拍的关系给了我们一个研究的起点，从这里，我们获得了一些视觉对应物（visual　analogues），对应于节奏的概念，包括强弱和节拍速率．视觉节拍同时改了我们一个操纵视频中的结构节奏：首先，我们识别视觉节拍，然后在时间刻度上规整这些拍子，使对齐它们向某一目标．加入我们可以识别关键的节拍，那么我们就可以把视频规整，对齐成任何我们选择的音乐．

1. １应用

计算视觉节奏有很多应用．我们主要关注那些与视频重定向（video retargeting）有关的应用，包含舞蹈的分析与合成．另外，这些应用也可以测试我们关于视觉节奏和舞蹈的基本假设，不至于白做了这些应用．

**舞蹈重定向：**通过弯曲时间刻度，现成舞蹈的视觉节拍可以与新的音乐对齐，这样我们就可以改变舞蹈表演者的配乐．假设有一个特殊的情况，即音频中，视觉节拍和音乐节拍已经对齐了，那么我们就可以很轻松的找到它们．利用这点，我们可以通过各种计算机视觉算法，测试我们关于视觉节拍和舞蹈的假设．

**舞蹈化**（Dancification）：我们关于视觉节拍的假设允许非舞蹈视频中包含视觉节拍，但是这些节拍不是按照某一可检测的节拍分布．如果我们可以单纯地通过视觉算法找出这些节拍，就可以把非舞蹈性质的视频转换成舞蹈性质的视频．我们称之为舞蹈化（Dancification）．我们也可以用这种方法，提升跟不上拍子的舞蹈，为舞蹈提供自动谐调（auto-tune）．

**意外的舞**（Accidental Dance）：我们可以把某一视觉节拍作为搜索标准，在大量的视频中找出一些像是跳舞的视频片段．只要找出这些片段，我们就可以按照视觉节拍讲这些片段穿插成任意长度的舞蹈视频．

视觉乐器：视觉节拍提供了短时的控制点，可以有更多的操纵视频的方式．比如，把规整的视觉节拍作为一个乐器的音符．（比如，录制音乐设备数字接口，或一个合成的表演）我们可以用这个乐器作为与音乐交互的接口，来编辑视频．

１．２　节拍的特点

我们从把节拍（包括音乐的和舞蹈的）的感知分解成不同的类型特点开始，从文学艺术观察中汲取[Bolton1894;Chionetal .;CowellandNicholls 1996; McPherson 2006; Vernallis 2004]，或者相关工作中，关于音频节拍追踪使用的启发式方法[Ellis 2007; Goto 2002; Grosche et al . 2010; Hu et al . 2017; Lerch 2012; McFee et al . 2015]．和舞蹈的计算分析[Brick and Boker 2011; chulLee and kwon Lee 2005; Dyaberi et al . 2006; Kim et al . 2003; Liaoet al . 2015; P. Chen et al . 2011] ．在第四节中，这些韵律学特点指导我们设计的寻找视觉节拍的启发式方法．在第五节中，是舞蹈相关的，解决时间规整视频的方法．

我们经常把一段音乐中，听众会用拍手或跺脚伴奏的时刻定义为音乐节拍．这种定义是采用某一音乐特征进行衡量的，不同的声音会通过不同的方式影响节奏的感知．大多数分析节奏的工作都是用这一特征来进行启发式地寻找音乐中的节拍的．这种方法通常由两个函数组成：一个是临时局部函数来衡量音乐的起始强度（onset strength），另一个函数衡量对一个指定节拍的符合程度，即起始强度分布的周期性．

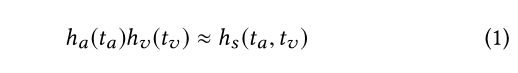
我们对于视觉节拍的定义暗含了一种相关的根植于舞蹈的感知的特征．我们假设这一特征也可以被分解成局部的和节奏的两个元素，从这两者中，我们可以为起始强度和拍子获取足够的视觉信息．需要注意的是，关于视觉节拍特征的和关于传统图片特征[Judd et al . 2009; Liu et al . 2009; Pritch et al . 2008]的局部信息是不同的，区别在于这是视觉动作的函数，它能够反映一些在时间上的局部事件．

我们把视觉和音乐节拍的节奏成分称为节奏特征（rhythmic saliency），把局部成分成为局部特征（local　saliency）

1. ３　同步特征

对舞蹈的感觉很大程度上受到伴奏的影响．这也是为什么舞蹈可以跟一首曲子很谐调，但是换一首就很不谐调．我们把讨论的这种协调性称之为同步特征，它衡量视觉和听觉事件之间联系的强弱．

我们用两个函数，表示听觉事件和表示视觉事件，如果它们的积满足：



则, 称它们是同步特征补函数(synchro-salient complements)

换句话说, 同步特征补函数是指在视觉与听觉信号上有高同步性的一对函数.

在第四节, 我们把针对视频局部和节奏特征的启发式算法设计成与音频节拍检测的启发式算法互为同步特征补函数. 这样我们可以将舞蹈化表达为与目标节奏特征的对齐化.

2 相关工作

计算编辑和视频制作是计算机图形学和计算机视觉中热门的领域, 有大量的工作从艺术中获得的灵感. 一些的焦点集中在基于电影中物体的自动化编辑任务. [Berthouzoz et al . 2012; Leake et al . 2017; Wang et al . 2008]. 另一些通过计算将视频转换成新的类型,如视觉化的东西, 摘要或总结. [Bai et al . 2012; Burg and Beck 2012; Chuang et al . 2005; Pritch et al . 2008; Schödlet al . 2000]. 像我们, Wang et al. [2014] and Bazin et al. [2016], 研究了根据物体函数的对视频的不均匀地时间规整, 将不同视频对齐的任务. Witkin and Popovic [1995], White et al. [2006], and Wanget al. [2006]研究了可以改变视频或动画的视觉风格的运动滤波算法, 该算法采用了在思路上与我们第五节中的时间规整算法相似的策略.

Davis et al. [2017; 2015a; 2015b; 2014]的工作和我们相似, 在于他们把音频处理中的概念运用到了视频动作中. 但是, 他们的工作集中在关键动作的频率上, 他们用这个从高速视频中提取声音. [Davis et al . 2014], 估计材质属性[Davis et al . 2017, 2015a], 为捕获到的物体建立可交互的物理模型[Davis et al. 2015b].

一些计算机图形学的工作把焦点放在音频和3D动画的同步上. [chul Lee and kwon Lee 2005; Kimet al . 2003; Langlois and James 2014], 的关于生成3D角色的舞蹈动画的工作, 与我们的工作相关. Kim et al. [2003] and Lee et al. [2005] 识别了在半周期人物动画中的短期控制点, 跟MIDI音乐同步, 生成动画舞蹈. 我们通过提取音乐和舞蹈之间,更关键的节奏结构的联系, 为更广泛的视频音频提供同步化.

与我们的工作更相关的是Liao et al. [2015] and Chen et al. [2011]的工作. 我们都把研究音乐作为操作视频的方式. 但是,他们的工作更多地被应用驱动, 并且主要关注于跟音乐有关的步子和片段的布置(pace or placement of clips relative to music). 与之相反, 我们关注于描绘音乐节奏和舞蹈之间的关系, 介绍用视觉拍子来指导更精确的和更动态的事件规整. 这些区别导致了看上去很不同的结果. Liao et al.产生了有时间流逝美觉的蒙太奇. Chen et al. 逐步地加速或减速视频来跟音乐对齐. 我们的结果是精确地合成舞蹈视频.

所有用音乐来驱动角色动画[chul Lee and kwon Lee 2005; Kim et al . 2003]和视频[Liaoet al . 2015; P. Chen et al . 2011]的相关工作都把舞蹈的特征分解为局部的和节奏的启发式函数. Kim et al. [2003]甚至用这个启发式函数定义一些与视觉节奏类似的东西. 但是, 他们的节奏特征的启发式函数衡量了错误的拍子的关系, 这限制了他们从有不同变化的拍子的原动画中识别视觉拍子的能力. 与他们相反, 我们的节奏特征函数扩展了Ellis [2007]用局部适应方法测量音频中拍子的动态规划方法, 这允许我们从不规则的视觉输入中找到视觉拍子. 我们还获取了一种弯曲特定舞蹈的视觉拍子的差值算法, 可以放大视频中的节奏特征.

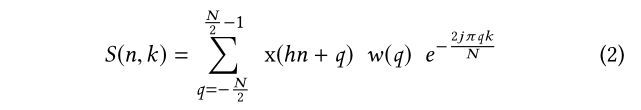
3. 量化音频中的节奏.

这里, 我们重新获取一种量化音频节奏的通用方法, 关注音频特征与(第四节中获取的)视觉特征之间联系的细节. 这一节中介绍的每一种算法都可以在Librosa python库中找到源代码[McFee et al . 2017, 2015], 更多潜在的理论由他们[Ellis 2007; Goto 2002; Groscheet al . 2010; Hu et al . 2017; Lerch 2012]解释. 音频特征和视频特征的视觉化描述在图2中.

从1D的音频信号开始, 我们重新获取了一种提取离散音频拍子的方法, 即时间上不同的一系列点. 我们的方法建立在’拍子是由音乐端点(?music onsets)在时间上的分布定义’的这一基础假设上.

3.1 语谱图和频谱流量

端点(onset)是指信号能量突然变大. 但是单独靠能量不足以探测出所有端点. 比如电子琴, 可以发出声调可以不断变化的持续的声音. 那么端点就是指音调发生变化, 而不是能量. 所以我们使用语谱图同时测量能量和音调, 计算音频信号的时间窗口的FFT.



产生一个复数矩阵S, 表示第n帧, 第k个点的频率值. W是窗口函数(比如汉明窗), h是步长(每两个窗口的之间的位移), 同时定义了n和t的关系. 的振幅近似于在n时刻,频率为k的能量x

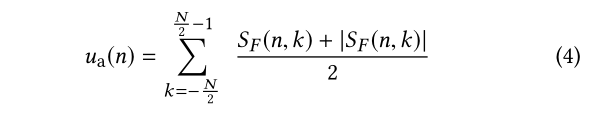
语谱图提供了频谱流量(spectral flux), 衡量了振幅在不同频率不同时间的变化, 可以用于寻找能量的端点[Böck and Gerhard Widmer 2013; Dixon 2006]:



3.2 端点包(?Onset Envelopes)

端点包(有时被称为新奇曲线(?novelty curves))是近似地衡量了每个时刻是一个端点的概率. 每一个端点大致上与某个频率上的频谱流量一致. 一个计算端点包的算法是把语谱图上各个频率的正的频谱流量相加, 得到一个非负的1D 时间-信号函数 [Böck

and Gerhard Widmer 2013; Dixon 2006]:



在3.5节中, 我们用从端点包中得到的端点强度作为检测音频节拍的局部特征的启发式函数.

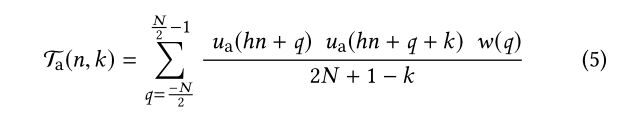
3.3 端点检测

中一个大的波动是端点的充分不必要的特点, 平滑的给端点检测带来更大的不确定性. 因此, 离散端点检测可以被公式化为取端点包中峰值. 一个简单的方法是取大于局部均值的一个局部最大值.

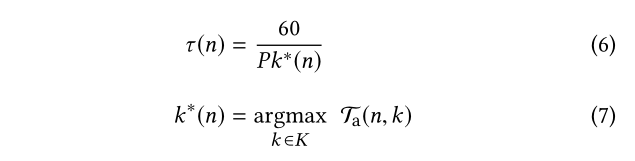
3.4 节拍速率和节拍速率图(Tempo and Tempograms)

节拍速率可以估计为端点包的自相关函数的峰. 这个峰表明了信号在某个周期上的自相似性. 节拍速率是在可计算的频率区间内(如,20-300bpm)最大的一个峰.

关于时间变化的拍可以被一个节拍速率图表示[Groscheet al . 2010], 它是从端点包中, 用与提取语谱图相似的方法,提取出来的. 区别是加窗FFT被替换成了非贝叶斯局部自相关函数(unbiased local autocorrelation), 方程式为:



其中, w(q)是一个5秒的窗. 的列是归一化的,展示了局部能量的影响, 是节拍速率关于时间的热点图. 第n帧上的节拍速率 t(n)衡量为每分钟节拍数(bpm),就是:



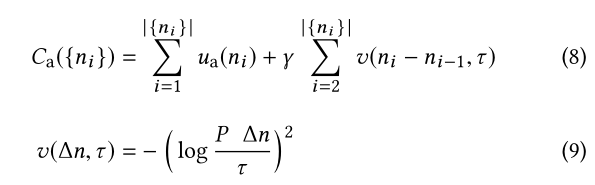
其中,P是帧区间, K是一系列可以计算的节拍速率的值(如,20-300bpm). 把搜索范围限制在到K有一个好处是在小的偏移上丢弃了假的峰, 其结果可以作为平稳信号上的局部自相似性, 在大的偏移上, 公式5的分母变成小的增强信号噪声.

另外, 根据Wiener-Khinchin理论, 提供了跟能量谱相似的信息. 然而, 谐波的能量谱的能量容易在分谐波的自相关性中显露. 我们早4.5节中讨论分谐波的好处.

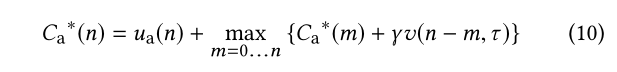
音频中的节奏结构可以视作能量在节拍延时间水平的分布. 它的谐波和次谐波如图2,3,7

3.5 节拍追踪

按我们在1.2节中讨论, 节拍追踪通常表现为对节拍特征的近似值的启发式优化. 这个近似值的局部成分会使拍点落在音频的端点上, 节奏成分会把拍点分布在一个不变的节拍速率上. 这导出了一个目标函数:



其中,是端点集, v是节拍速的成对对象处罚偏差(pairwise objective penalizing deviation). 节拍次数的最优序列可以由根据递归关系的动态规划得出:



其中, a控制相关权重.每个都可以得到一个节拍序列. 分值最大的序列作为结果返回.

4 视频中节奏的量化

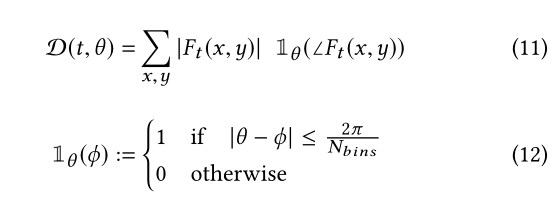
在这节中我们获取对应于听觉的视觉的特征. 图2将这些对应的特征与音乐特征可视化展现出来.

我们从为视频的局部特征选择一个启发式函数开始. 总的来说, 我们希望这个函数完全与音频中的对应. 在第三节中, 是端点强度.端点强度在脉冲和脉冲响应的端点处最大化. 在物理世界, 这些会与运动物体的冲击和共振一致, 通常导致运动物体突然的减速. 我们用这些视觉上突然地减速作为我们对局部特征的启发式函数, 我们称之为视觉冲击点(visible impact).

我们的输入是常规的视频, 使用Bouguet [yves Bouguet 2000]的方法, 从视频的每一帧到下一帧,计算光流.

4.1 方向图(Directograms)

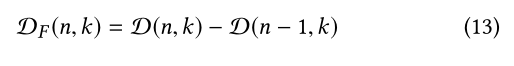
在3.1节中,我们用语谱图S将音量分解成各个频率. 现在, 我们计算2D矩阵, 我们称为方向图, 它将运动分解为各个角度. D的每一列都是每一帧t的光流的角度的加权直方图:



这里, 是一个指示函数, 用来将流向量分解为N个方向.

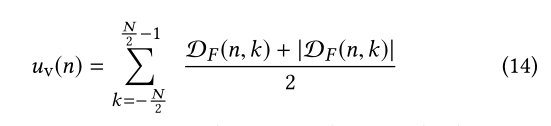
某些视频编码在视频中会有重复的帧, 导致D中有空白的列. 我们给D加一个3x3的中值滤波器, 来解决这些列. 注意, 核的每一维都有必要为曲线运动做出解释.

现在, 我们的方向图看起来跟我们在3.1节中语谱图的振幅很像: 一个标量矩阵且x轴对应时间. 这样, 通过公式13,我们可以计算每个方向的减速度作为方向图的频谱流量(对应于音频的频谱流量):



4.2 冲击点集

等式13给了我们一个矩阵, 跟频谱流量形式一样. 然后我们计算端点集的对应物,冲击点集, 把列上所以正的值相加得到.



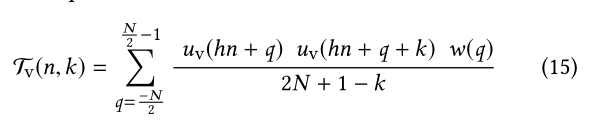
这给了我们一个冲击点集, 它的形式跟端点集一样. 在视频的剪辑处, 会有一个超级大的峰, 为了去掉它, 我们把值中最高的2%去掉, 然后归一化.

4.3 冲击检测.

为了检测离散的可视冲击. 我们先计算局部均值, 用0.1s的窗口. 然后计算局部最大值, 用0.15s的窗.然后,我们定义冲击点是局部最大值超过局部均值达到至少全局最大值的10%的点.

4.4 视觉节拍速率图

视觉节拍速率图通过把等式5中的替换成得到:

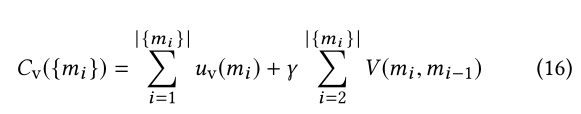


配合音频的节拍速率图, 我们可以看到视觉节拍速率图中的节奏结构是一条水平线, 其下是谐波和分谐波. 图2,3,7显示了舞蹈视频中音频和视频的节拍速率图中有一致的结构, 这说明, 在某种条件下节拍速率图是同步特征的互补.

4.5 可是节拍追踪.

对于舞蹈视频, 我们可以用3.5节中同样的算法得到视觉节拍. 然而, 大多数应用都是用视觉节拍点作为时间规整的控制点. 这种情况下, 我们选择视觉节拍的标准就有些不同了, 选择的质量会在输出中量化. 为了衡量这个, 我们必须考虑局部特征和节奏特征规整.

我们的局部特征是视觉冲击, 它是不连续的减速的估计. 我们先保证时间规整不会在我们的输出中制造错的冲击点, 这在对连续的动作使用不连续的比率的时间规整是会发生. 为了避免这种错的冲击点, 我们把视觉节拍点限制在的局部极值上. 然后我们强制时间规整的比率在除了视觉节拍点之外的所有地方是连续的, 这保证了不会产生新的视觉冲击点. 只考虑检测到的冲击点, 我们现在定义一个跟等式8相似的目标函数:



这里的目标V有多重选择.

在音频中, 我们用目标去限制信号主节拍的变动. 在重定向的应用中, 我们通常假定不存在这样一个主节拍. 我们的工作是创造一个主节拍. 我们可以限制信号的主节拍的变动作为一种把时间规整比率调整到接近1的方式.

另外一个选择是设V=0, 从而把所有冲击点都作为视觉节拍. 当所有视频中的动作都很大时, 这个策略会很有效. 并且这保证了所有的视觉冲击点都对应节拍点. 但是, 这会使结果与4.3节中描述的窗口和阈值的大小有很大的相关性, 尤其是有频繁的小的动作时.