ISSN 1000-0054 清华大学学报（自然科学版）2008 年 第 49 卷 第 1 期 w30
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2008 Vol. 49, No. 1 http://qhxbw.chinajournal.net.cn

基于事件序列匹配的多摄像机视频同步

黄飞跃[1,2]， 徐光祐[1,2]

（1. 清华大学 计算机科学与技术系，普适计算教育部重点实验室，北京 100084；
2. 清华信息科学与技术国家实验室，北京 100084）

摘 要：在多摄像机应用中，对各摄像机的视频进行同步是一个常用的基本步骤。大多数现有方法都依赖于特征点的检测和对应，往往复杂而且难以通用。该文致力于简单快速视频同步方法的研究，通过对常见同步问题的描述和分析，提出了一套基于事件序列匹配的视频同步框架——利用事件序列相关性来帮助寻找两段视频序列中的对应帧，通过全局投票的方法寻求最优帧偏移量。在动作识别和会议建档系统场景中的示例实验中都得到了正确的结果。误差分析表明，该方法的容错能力强，在较大参数范围内都可以得到正确解。该文提出的方法简单易用，具有很好的实用价值。

关键词：视频同步；多摄像机；事件序列

中图分类号：TP 391.4 **文献标识码**：A

文章编号：1000-0054(2009)01-0118-05

Event sequence matching based multi-camera video synchronization

HUANG Feiyue[1,2], XU Guangyou[1,2]

(1. Key Laboratory of Pervasive Computing of Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China)

Abstract: Video synchronization is a common issue in multi camera applications with most approaches depending on feature detection and point correspondence which are complicated and difficult to use. This paper provides a simple, fast method for video synchronization based on event sequence matching. This method uses the relationships between event sequences to match frames between two video sequences and seeks global optimum frame offset using statistical methods. Correct results were obtained in tests of action recognition and spot archive scenarios. An error analysis shows that this method is tolerant to errors when the algorithm parameters are properly chosen so the system is convenient to use in practical cases.

Key words: video synchronization; multi cameras; event sequence

在计算机视觉尤其是运动物体的检测和识别研究中，对多摄像机协同的需求日益普及。多摄像机之间的标定[1]和同步是多摄像机协同应用中两个关键的问题。多摄像机间的标定是一个比较成熟的问题，而摄像机间的视频同步则往往被人们忽视。在实际的多摄像机应用中，为融合多摄像机的信息，各摄像机的视频同步是一个必须的步骤。目前通常采用人工的方法，比如通过在多路视频采集时加入特殊标记或通过开关灯、挥手等方法来标记，然后人工识别和记录各个摄像机之间的偏移量。

当前对视频自动同步的研究还没有引起足够的重视。已有的方法主要有以下几种：利用双摄像机之间的成像几何约束[2]、利用双摄像机对应点的秩约束[3-4]、利用时空兴趣点的相关性直接同步时间偏移量[5]、整体匹配时空对应点[6]等。这些方法大多都利用摄像机之间的特征点对应来实现视频中的单组对应帧匹配，往往忽视了时间轴上的序列相关性信息。由于这些方法是以实现精度要求高、实施困难的对应特征点检测和跟踪为基础，这大大地限制了它们的实用价值。文[7]中采用了基于表观时序相关性的视频同步方法，这种方法虽然避免了对应特征点的检测，但是它利用全局搜索同时寻找最优区域和最佳帧匹配两种最优值，因此有时会得到完全错误的结果。比如当图像前景突变或者摄像机偶尔振动的时候，突变的一两帧图像引起的误差就会导致完全错误的偏移量求解。

本文利用双摄像机视频序列中的视觉信息事件序列来帮助进行对应帧匹配和对准，从而绕过了摄像机特征点对应这一较困难的问题，无须逐一寻找对应帧。利用时间轴上两组事件序列的相关性，采用对偏移量投票的方法来求全局统计最优解。

收稿日期：2007-11-20
基金项目：国家自然科学基金资助项目（60673189，60433030）
作者简介：黄飞跃（1979—），男（汉），江苏，博士研究生。
通讯联系人：徐光祐，教授，E-mail: xgy-dcs@tsinghua.edu.cn

1　摄像机视频同步问题分析

在视频采集系统中，通常把摄像机和视频采集卡结合使用，记摄像机为C_1和C_2。采集过程为：摄像机间隔固定的采样周期进行采样，在某一时刻采集卡启动并开始采集视频。记2个摄像机对应的采样周期为T_1和T_2，2个采集卡开始采样的时刻为t_{s1}和t_{s2}。那么采集到的视频序列中的第N帧视频实际对应的采样时刻分别为：

$$t_1 = t_{s1} + N T_1, \quad t_2 = t_{s2} + N T_2. \tag{1}$$

摄像机的视频同步问题就是要在2个摄像机采集的视频帧序列中得到同一时刻采样的图像帧的对应关系。也就是已知摄像机C_1中的帧号n_1，求同一时刻摄像机C_2采集序列中的帧号n_2，或者反之。如式(2)所示。

$$t = t_{s1} + n_1 T_1 = t_{s2} + n_2 T_2. \tag{2}$$

由式(2)又可以推出

$$n_2 = n_1 T_1 / T_2 + (t_{s1} - t_{s2}) / T_2. \tag{3}$$

在实际的多摄像机应用中，通常利用相同采样周期的摄像机(一般是同一种型号的摄像机)，那么$T_1 = T_2 = T$，于是，

$$n_2 = n_1 + (t_{s1} - t_{s2}) / T. \tag{4}$$

按照式(5)定义帧偏移量

$$F = (t_{s1} - t_{s2}) / T. \tag{5}$$

此时，摄像机的同步问题就变成只需求解2个摄像机视频序列的帧偏移量F。一般情况下，它不一定是整数。不过可以通过在摄像机之间接入外同步信号来简化这个问题。此时各个摄像机始终都在相同的时刻点采样，这时F必然是整数。理论上只需得到两个摄像机之间的一组对应帧n_1和n_2。那么

$$F = n_2 - n_1. \tag{6}$$

在实际应用中，由于无法保证n_1和n_2会严格对应匹配，通常需要得到尽量多的对应匹配帧，然后通过大量对应匹配帧来求解最优偏移量。

2　基于事件序列匹配的投票方法

在上节的分析中，可知道摄像机同步问题可以简化成2个视频序列帧偏移量的求解，而视频同步的关键就是查找2个视频中的对应匹配帧。然而因为2个摄像机的摆放位置不同，没有固定的几何位置约束，直接逐一查找对应匹配帧还是比较困难的。那么是否可以避免逐一寻找对应帧呢？作者尝试考虑利用序列相关性来寻找偏移量。

对于一段待同步的视频序列，定义某时刻客观发生的某种事件E，它和主观观测无关，比如前景中人的出现和消失、开关灯等。如果事件E在t时刻发生，称t时刻摄像机采集的图像帧为事件帧。定义事件函数$E(n)$如下：

$$E(n) = \begin{cases} 1, & \text{第 } n \text{ 帧是事件帧;} \\ 0, & \text{第 } n \text{ 帧不是事件帧} \end{cases} \tag{7}$$

$E(n)$函数上所有满足$E(n) = 1$的n组成事件帧集合N。对于视频序列对1和2，客观上存在符合定义的事件帧集合N_1和N_2，称之为真实事件帧集合，它们的真实事件函数曲线E_1和E_2是完全相关的，仅仅差一个偏移。即

$$E_2(n) = E_1(n + F). \tag{8}$$

这样，不需要从单帧匹配入手寻找对应关系。只需利用事件帧序列，由2组函数曲线的相关性即可求解偏移量。由于函数是离散并且阶跃的，因此可以很方便地衡量两者的匹配度。作者仅用$E(n) = 1$即事件发生的时刻来衡量匹配度。对于E_1和E_2，定义它们关于偏移量F的匹配度为

$$M = \sum_{\substack{\text{满足条件} \\ \text{的} n}} E_2(n), \quad E_2(n) = E_1(n + F) = 1. \tag{9}$$

只需对可能范围内的所有偏移量进行遍历，那么使M最大的偏移量就是最优的解，即

$$F = \arg\max_F M. \tag{10}$$

实用中，只能通过观测来验证某个事件的发生。由于2个摄像机的相对关系未知，通常只能选用单视图的观测特征来检测，比如人的出现事件是通过对图像前景的检测来实现。可以通过事件的观测特征，来检测相应的事件帧。通过观测特征检测得到的事件帧集合称之为观测事件帧集。观测事件和真实事件存在偏差，但由于无法得到真实事件，只能用观测事件来代表真实事件，通过它们的相关性来求解最佳偏移量。根据上述分析，给出偏移量的求解方法——基于事件序列匹配的投票算法。

1) 定义便于检测的事件及其观测特征。

2) 在2个摄像机帧序列中，检测出观测事件帧，把对应帧号加入各自的观测事件集合。

3) 在一个范围内，遍历各个偏移量，计算两个事件帧集合中满足式(6)的匹配帧对数。每存在一对匹配帧，相当于对此偏移量投了一票。

4) 选择得到最大投票M即匹配帧数对多的偏移量作为最优解。

适用于本算法的事件通常应该符合如下要求：a) 事件应便于在单视图中检测，出现漏检测和误检

测的概率比较低; b) 由于摄像机可能在任意位置, 事件的观测特征应该尽量和摄像机位置无关; c) 事件发生的频率适中而且非周期性, 这样有助于利用相关性进行对准。

通常可以根据实际应用场景, 选择动态特征或者语义特征对应的事件, 这样的事件观测特征更容易具有摄像机位置无关性, 比如光照变化、物体运动拐点、前景目标变化、背景切换和变化等。

3 示例应用场景实验

3.1 动作识别

在多摄像机的人体动作识别[8-9]的应用场景中, 需要融合2台摄像机的信息以获取与人体绕垂直轴转角无关的特征, 因此视频信息的同步是实现此方法的关键。图1是应用场景的实际数据示例, 本场景中摆放了2个光轴夹角接近90°的摄像机。

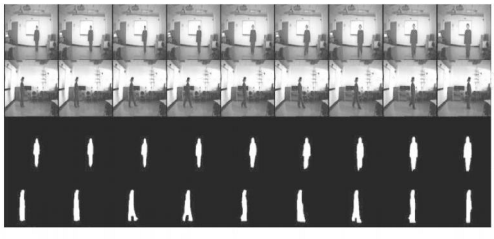

图1 人体动作识别应用中的数据示例

1) 人体静止状态事件。

在动作识别中, 最明显的特征是人体运动信息, 因此使用人体运动信息来定义事件。首先选择人体静止事件。定义人体静止时刻就会产生静止状态事件, 采用人体运动区域所占比例阈值化作为观测特征。理想情况下运动区域为0才是静止事件; 但实际中, 当运动区域小于某个阈值就认为它是静止事件。具体做法如下: 利用PFinder方法[10]提取人体区域, 然后把当前帧和上一帧的人体区域做图像相减, 计算出每帧中人体运动区域。运动区域的大小反映了当前时刻人体运动幅度。如果人体静止, 那么运动区域面积理论上为0; 而如果人体运动幅度较大, 那么运动区域面积也相应较大。图2是2个摄像机中人体区域面积占图像面积的百分比$F(n)$和运动区域面积占图像面积的百分比$P(n)$对应的时序曲线图。在两个摄像机对应视频中各取500帧作为实验数据, 图中只显示了前300帧。

通过对运动区域进行阈值化来检测静止事件帧, 定义静止事件的函数$E(n)$为:

$$E(n)=\begin{cases}1, & P(n)<T_{\min};\\ 0, & P(n)>T_{\min},\end{cases} \tag{11}$$

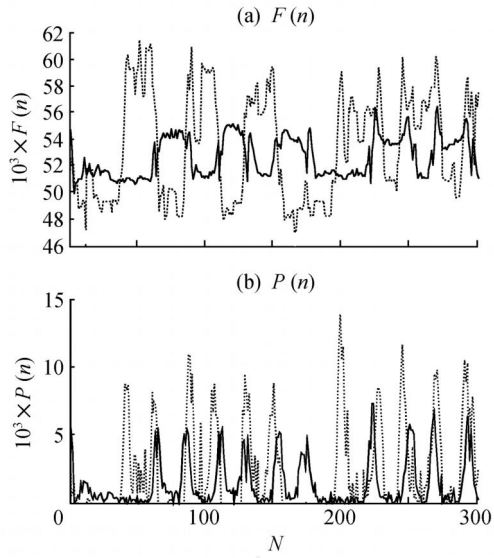

图2 时序曲线

阈值$T_{\min}$的选择方法如下: 统计或者指定人体区域的面积百分比A, 阈值$T_{\min}=A\alpha$。图3a和图3b分别是$\alpha=0.01$时的$E_1(n)$和$E_2(n)$函数图(使用柱状图表示)。由于本方法基于统计投票, 因此对于误差和阈值的选取具有较好的容忍度。图4显示$\alpha=0.01$(实线)和$\alpha=0.02$(虚线)时不同偏移量下的匹配投票值, 可以看到两者都是在偏移值$F=-23$处取得了最大匹配值。

图3 $\alpha=0.01$时$E(n)$函数

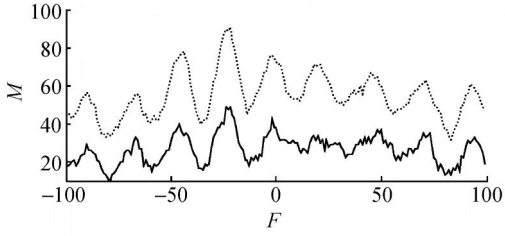

图4 不同偏移量对应的点集匹配数($\alpha=0.01, 0.02$)

2) 人体较大运动事件。

也可以定义其他事件进行相关性匹配, 比如人体较大运动事件。和人体静止类似, 采用对运动区域阈值化的方法来表示观测事件, 定义较大运动事件的函数$E(n)$为:

$$E(n)=\begin{cases}1, & P(n)>T_{\max};\\ 0, & P(n)<T_{\max},\end{cases} \tag{12}$$

阈值$T_{max}=A\beta$。图5a、b分别是$\beta=0.01$时的$E_1(n)$和$E_2(n)$曲线。图6是$\beta=0.1$(实线)和$\beta=0.15$(虚线)时不同偏移量下点集匹配结果, 最优解和人体静止事件得到的结果都完全一致。

图5 $\beta=0.01$时$E(n)$函数

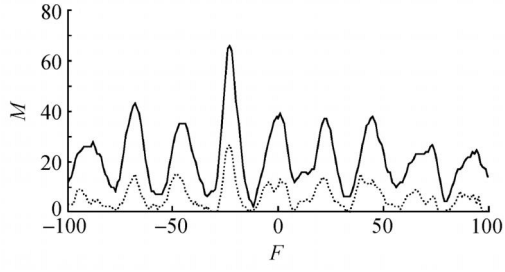

图6 不同偏移量对应的点集匹配数($\beta=0.1, 0.15$)

3.2 会议建档系统

会议建档系统通过在会议室中部署多台摄像机, 基于动态上下文场景融合多摄像机信息进行事件分析[11]。由于需要融合多个摄像机的信息, 因此同样需要进行视频同步。如图7所示。

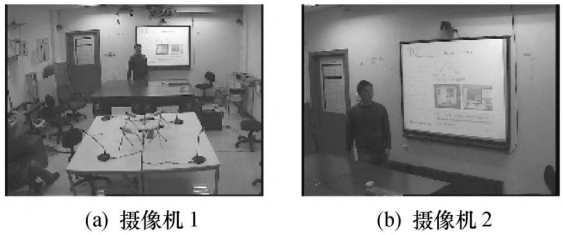

图7 会议建档系统场景示例

在该场景中, 存在多种可以用于同步的事件, 比如举手、人员走动、起立坐下、开关灯等等。这里, 采用幻灯片切换事件作为示范。利用对幻灯片背景区域的像素亮度变化的统计结果阈值化作为切换事件的观测特征。定义了2种统计方式进行比较, 参见式(13), 其中$I_t(p)$表示t帧p点像素的亮度值。对统计结果$P(n)$进行阈值化就可以得到事件函数$E(n)$。

$$\begin{cases} P_1(n)=\mathrm{Avg}(I_t(p)-I_{t-1}(p)), \\ \quad p\ \text{在ROI区域}; \\ P_2(n)=\mathrm{Avg}(I_t(p)-I_{t-1}(p))^2, \\ \quad p\ \text{在ROI区域} \end{cases} \tag{13}$$

图8是2个摄像机的$P_1(n)$函数曲线图, 其中较大取值处对应了幻灯片切换事件的发生。图9是分别利用P_1和P_2进行匹配投票得到的匹配曲线。本实验中, 演讲者遮挡、幻灯片变化微小等都会影响切换事件的检测; 而不同的统计方法和阈值选取也会影响事件检测。但是基于相关性匹配的投票同样都可以得到正确解, 说明了这种算法的鲁棒性。

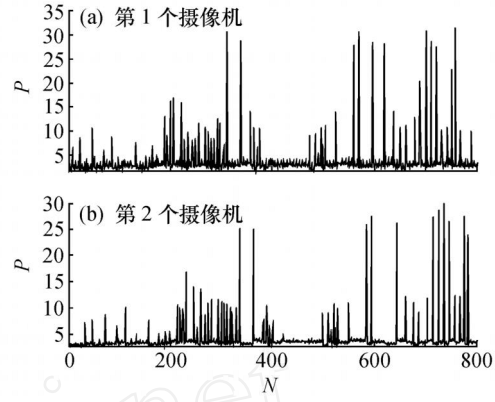

图8 2个摄像机的幻灯片区域像素亮度变化曲线P(n)

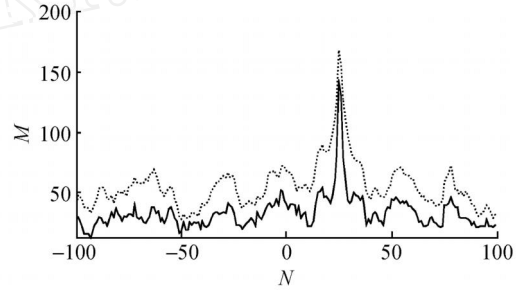

图9 不同偏移量对应的点集匹配数(P_1和P_2两种统计)

4 误差分析

观测事件帧集和真实事件帧集合通常存在一定的差别, 包括事件帧被漏检测或者非事件帧被误检测为事件帧。漏检测会导致真实偏移量的投票得票数减少, 而误检测则会导致某个非真实偏移量的得票数增加。不过本算法中一次漏检测或误检测只会引起偏移量的一票误投, 不会对解有突变影响。

4.1 理论分析

计序列帧总长为N, 其中真实事件帧数为R, 2个序列中的漏检测帧数为A_1和A_2, 误检测数为B_1和B_2。假设误差和事件是均匀分布, 暂不考虑它们的分布特性。那么投票中, 各偏移量平均得票数为

$$M_1=\{(R-A_1+B_1)(R-A_2+B_2)-[R-\max(A_1,A_2)]\}/N. \tag{14}$$

而真实偏移量得到的票数为

$$M_2=R-\max(A_1,A_2)+M_1. \tag{15}$$

定义事件发生的频率为x, 即$R=xN$, 同时误检测和漏检测发生率为y ($0<x,y<1$), 那么可以

推出

$$\frac{M_2}{M_1} \approx 1 + \frac{1}{x(1-y)}. \tag{16}$$

偏移量的投票曲线类似图4、6、9所示,存在若干波峰波谷。理论上,M_2就是最高波峰值,而M_1是曲线的均值。事件分布和误差会导致投票曲线在均值线上下扰动。若M_2和M_1相比很大,则说明最高峰值比平均线要大很多,说明可容忍更多扰动。通常情况下,选择合适的事件,使得事件发生概率和误差率都适中而不太高,此时M_2/M_1值比1大很多。如假设$x=0.15$, $y=0.1$, 此时$M_2/M_1=6.5$。

以上只是考虑均匀分布,实际上由于一般事件和误差发生的不规则性和不可预期性,无法推导出一般意义的误差公式。下面采用实验进行误差的统计分析。

4.2 实验统计

本算法取最高的波峰处作为最优解。误检测或者漏检测会导致投票曲线有所变形,表现有2种形式,一种是最高波峰的位置做了小幅偏移,而另一种则是非最高峰变成了最高峰。第1种变形导致了偏移量求解会有小幅偏差,这对应算法的精确性;而第2种变形则是求解的跳变错误,求解的偏移量和真值可能有较大误差,这对应了算法的鲁棒性。

利用较大数量的实验数据对误差进行统计分析。用基于双摄像机的人体动作识别文[9]中的数据,样本有7人,每人3组视频对,共有21组视频对,每对视频包含约2 000帧。所有视频对的实际帧偏移量分布在[-43, 57],绝对值平均约23。实验分别选用500、1 000、1 500帧的序列对,用3.1和3.2描述的2种事件,在[-100, 100]偏移量范围进行遍历投票,最后得出的统计结果见表1。每种事件有3个子列,是对这些视频求解结果的3种统计:A——最高峰正确时,偏移量真值和解的平均差值;B——非最高峰被当成最高即解存在跳变错误的个数;C——最高峰无误时,最高峰和次高峰的平均比值。

其中,A反映了本算法的精确性,B和C则反映了算法的鲁棒性。B越小,说明突变性错误很少。而C越大则表明了最高峰不易跳变,对事件的误检测和漏检测的容忍度越大。实验结果中,跳变错误都很少,平均偏差也较小,说明在选择了合适的事件特征后,算法的鲁棒性和精确度都很好。

表1 同步实验结果统计

帧长	静止			较大运动		
	A	*B*	*C*	*A*	*B*	*C*
500	0.52	0	1.27	0.67	1	1.18
1 000	0.43	1	1.35	0.76	0	1.22
1 500	0.71	0	1.46	0.86	1	1.30

4 结 论

本文介绍了利用摄像机视频序列中的事件匹配来进行对准。通过选择和定义合适的事件,可以实现精确而鲁棒的摄像机视频自动同步。理论分析和实验表明,该方法简单易用,对误差的容忍度很好,具有很好的实用性。

参考文献 (References)

[1] ZHANG Zhengyou. Flexible camera calibration by viewing a plane from unknown orientations [C]//Proc 7th Int Conf Computer Vision. Kerkyra, 1999: 666-673.

[2] ZHOU Chunxiao, TAO Hai. Depth computation of dynamic scenes using unsynchronized video streams [C]//Proc Int Conf Computer Vision, CVPR '03. 2003: II 351-358.

[3] Tresadern P, Reid I. Synchronizing image sequences of non-rigid objects [C]//Proc IEEE Conf Computer Vision and Pattern Recognition. Washington, 2004.

[4] Rao C, Gritai A, Shah M. View-invariant alignment and matching of video sequences [C]//Proc IEEE Int Conf Computer Vision. 2003: 939-945.

[5] Yan J, Pollefeys M. Video synchronization via space-time interest point distribution [C]//Advanced Concepts for Intelligent Vision Systems. 2004.

[6] Caspi Y, Irani M. Spatio-temporal alignment of sequences [J]. *IEEE Trans on PattAnalMach Intelli*, **24**(11): 1409-1424.

[7] Ushizaki M, Okatani T, Deguchi K. Video synchronization based on co-occurrence of appearance changes in video sequences [C]//Proc 18th Int Conf Patt Recog. 2006.

[8] HUANG Feiyue, DI Huijun, XU Guangyou. Viewpoint insensitive posture representation for action recognition [C]//Proc IV Conf Articulated Motion and Deformable Objects. Mallorca, 2006.

[9] Huang F Y, Xu G Y. Viewpoint insensitive action recognition using envelop shape [C]//Proc 8th Asian Conf Computer Vision. Tokyo, 2007.

[10] Wren C, Azarbayejani A, Darrell T, et al. Pfinder: Real-time tracking of the human body [J]. *IEEE Trans on PattAnalMach Intelli*, 1997, **19**(7): 780-785.

[11] DAI Peng, DI Huijun, DONG Ligeng, et al. Group interaction analysis in dynamic context [J]. *IEEE Trans on Syst, Man, and Cyber, Part B*: 2008, **38**(1): 275-282.