基于视频的人机交互中动作在线发现与时域分割

沈 晴 班晓娟 常 征 郭 靖
(北京科技大学计算机与通信工程学院 北京 100083)

摘 要 时域分割问题是计算机视觉领域长期存在的问题.尤其在基于视频的人机交互过程中,动作的发现和分割被要求在线完成.对此,文中提出了一种基于模板的识别框架.对于不断产生新数据的在线视频序列,该方法可根据已知的信息,及时在线发现和分割已完成动作.该方法主要过程为:首先,通过基于鞅过程的算法提取关键帧,然后沿着关键帧对前序帧进行回退式遍历,接着通过构建动作历史图像来描述动作信息,最后通过计算相对于模板动作的包含率和七阶不变矩相似度实现动作的最优分割.在 IXMAS 数据集上进行实验,该方法的平均动作发现率达到了 88%,准确率达到了 75.9%,在基于深度数据的在线实验中,该方法获得了 82%的平均动作发现率.

关键词 动作时域分割;关键帧;动作历史图像;人机交互
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2015.02477

On-Line Detection and Temporal Segmentation of Actions in Video Based Human-Computer Interaction

SHEN Qing BAN Xiao-Juan CHANG Zheng GUO Jing
(*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083)

Abstract Temporal segmentation is a long-standing issue in computer vision. Particularly, in the vision based human-computer interaction, the on-line implementation and real-time performance is required. In this paper, a template based approach is proposed. Given a continuous video sequence, a completed action can be detected and segmented on-line, based on the previous movement information. In this approach, characteristic frames are extracted firstly by a martingale framework, followed by the back tracking along the characteristic frames, and then the motion history images are built to represent the motion information, which are used in the final segmentation via the calculation of the action inclusion rate and the likelihood of the seven invariant moments with the templates. The experiment on the IXMAS database got an average action detection rate at 88% and an accuracy at 75.9%. In the experiment on the on-line depth data, the detection rate reached to 84.8%.

Keywords action temporal segmentation; characteristic frames; motion history image; human-computer interaction

1 引 言

近年来,对视频中人体行为的分析,在计算机视觉领域一直是热门的研究方向.在识别人体动作类别方面,大多数研究集中在特征值提取、分类器训练等方面,通过有效排除背景干扰,提高视角变化适应性等途径,提高识别精度.这些研究所针对的对象是

收稿日期:2014-07-06;最终修改稿收到日期:2015-06-27.本课题得到国家自然科学基金(61272357,61300074)、教育部新世纪人才计划(NCET-10-0221)和博士后科学基金(20100480199)资助.**沈 晴**,男,1988 年生,博士研究生,主要研究方向为人机交互、机器学习.E-mail: shenqingcc222333@gmail.com. **班晓娟**,女,1970 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为人机交互与三维可视化技术. **常 征**,男,1986 年生,博士研究生,主要研究方向为计算机视觉、机器学习. **郭 靖**,男,1990 年生,硕士研究生,主要研究方向为计算机视觉、智能系统.

已经人为分割的视频段,动作数据拥有明确的起点和终点.

但是在实际应用中,人体动作信息是持续不间断的,各个动作(包括静态的姿势有机地结合在一起)共同组成了人类肢体语言表达.有效地发现和提取运动中的有意义动作,是通过肢体语言进行交互的基础和前提.如果不能排除人为干预,对视频进行自动时域分割和自动动作发现,单纯对动作进行识别是缺乏广泛应用价值的.

动作的时域分割的难点在于,缺乏有效的关于各帧之间人体运动相关性信息,尤其是缺乏可量化的动作分割的先验知识.所以只有同时针对动作的时域和空域特征,来准确地直接对视频进行分割并识别类型,成为当前动作识别技术发展中的重要方向.

随着一系列新型交互设备(如 Google glass、Kinect、Leap Motion 等)的出现,针对基于视觉的人机交互(HCI)中人类行为的分析,成为了动作识别技术的应用热点.与传统的视频分析相比,人机交互中的动作进行识别有着更为严格的要求:当用户需要用复杂动作以实现具体指令时(如挥舞手臂、拍手、转身等),计算机需要在缺乏后续内容的情况下,在线判断一个指令是否完成并进行识别、反馈.这一系列过程必须在用户进行下一个动作之前完成,计算机在用户指令发出后过久才给予响应,将使用户等待过久,严重影响操作的连贯性和交互效率.

当前基于手势的人机交互系统,许多只根据肢体部位的位置和角度特征来判断指令的输入,这种方法无法实现复杂动作指令的输入,如挥挥手关闭当前应用、打个响指确认选择、拍拍手开始一段音乐等等.所以,一个交互系统必须能在对用户动作的持续监测中,将用户的连续动作及时发现,并在线分割成可识别的单个动作.基于内容的在线视频分析系统,事件记录系统和安全警报系统也存在同样的要求.

对连续动作的在线分割识别,计算机视觉时域分割领域的主要观点主要集中在通过分析视频的统计特征值来表示视频各帧的时空关系,然后由经过训练的模板或分类器来确定具体的动作类型和分割位置.在离线分析的应用中,例如监控视频检索等,利用运动兴趣点[1]或光流信息[2]表示时空特征,并基于 HMM[3]框架或利用 SVM[4]等分类器确定分割点的方法,可以得到较高的识别准确率,

但是这些方法对视频的每一帧都需要识别,系统运行效率低.基于关键帧的姿势匹配和识别方法有效解决了这一问题,其中 Lu 等人[5]提出了一种基于关键帧对的方法,在连续动作分割和识别上有很好的结果.但是正如之前所述,以上这些方法在 HCI 应用中仍然有如下几个方面的缺陷:

(1)动作分割的实现,是在所有动作的完整视频流被输入后,系统通过对视频流求全局最优解而获得的.但是 HCI 系统是通过摄像机对用户行为进行不间断检测,视频流是持续且增量式产生的,没有固定的视频尾端,每一次数据的输入都可看作视频的一个结尾,却并不意味着当前已经有动作完整发生,同时分割结果需要在每个动作发生后立刻产生而不能或者极少依赖后续数据,所以在这种情况下,分割位置只能通过寻找当前已知视频序列的局部最优解得到.

(2)动态过程的时间相关性信息缺失严重.使用有限数量的特殊帧来表示一个动态过程,不可避免地发生信息缺失.在线识别系统中,不同动作的相似静态姿势的出现时,由于缺乏连贯的动态时空信息,很容易产生不当分割.

在 HCI 系统中,动作时域分割的应用问题可通过以下两个方法来解决(如图 1):

(1)每当有新数据帧被输入,就对待分割的前序序列进行回退式扫描并识别,寻找每一次回退后的视频段中的分割参考,并在这些参考位置中找到局部分割的最优解,从而判断待分割序列中是否有动作产生,以保证当前动作的及时发现.

(2)回退扫描的动作段的动态信息,由历史动作图像 MHI[6]表示.MHI 通过灰度值变化,体现了各帧之间的时域相关性,同时不同灰度值的人体轮廓保存了前序每一帧的运动信息和当前帧的姿势信息.

但是,这种回退式扫描方法将产生巨大的计算量,无法满足 HCI 的实时性要求,对此系统需要进行如下改进(如图 2):

(1)视频关键帧将被提取,识别过程可沿着前序关键帧位置进行跳跃式回溯.由于前后两帧的运动信息改变程度不大,所以在只增加了少数帧数据的情况下,回溯计算的结果差异不大,只有当人体位置发生明显改变后,回溯计算才可能产生有明显区别的分割参考位置,从而才有提取最优解过程的必要性.所以通过分析视频各帧的内容改变程度,提取关键帧作为回溯运算开始的标志,可以很大程度上提高运算的实时性.

(2)在每次回退后,并不对此段动作的 MHI 进行识别,只对此 MHI 分析相对于每个模板的动作

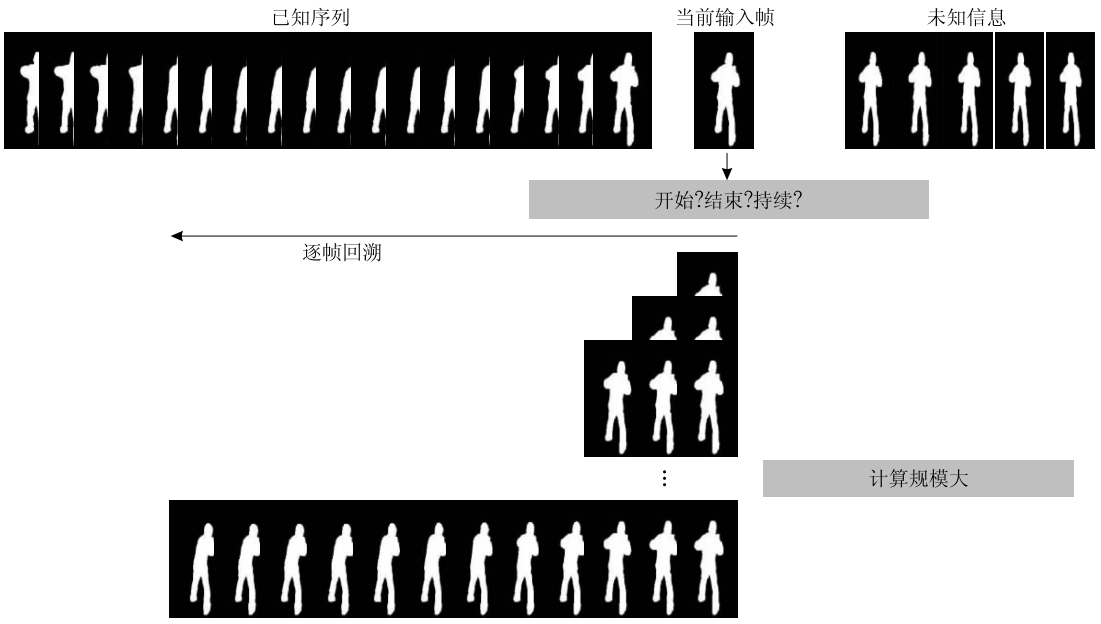

图 1　从增量式视频中回溯式寻找分割的局部最优解

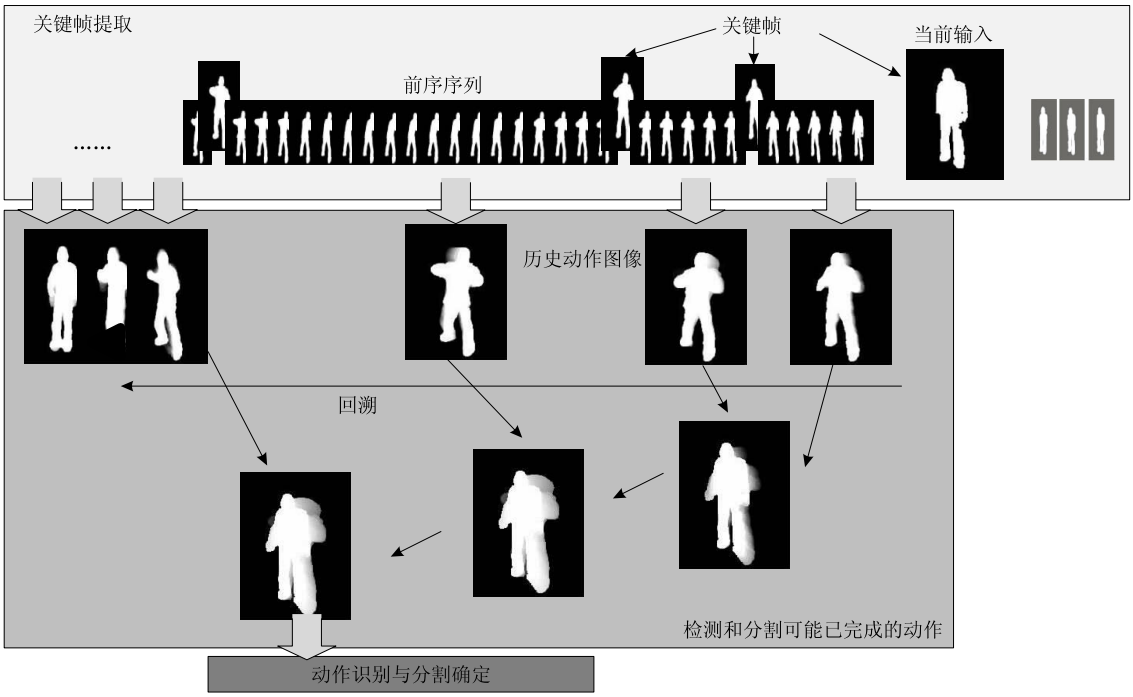

图 2　沿着关键帧进行跳跃式回溯检测可能已完成的动作

包含率，并初步判别是否已经满足一个动作出现的标准. 在包含率超过一定阈值时，才对其进行识别. 包含率的计算规模远低于本文识别过程所使用的基于 7 阶距的算法.

由于采用了以上改进，回退式扫描算法所产生的大规模运算损耗被大大减小，识别速度更快，从而使得 HCI 对动作自动发现的两个方面要求——不间断检测和运算实时性都得到了满足.

本文第 2 节简述时域分割领域的相关研究；第 3 节陈述关键帧提取算法；第 4 节介绍表示动作方法以及动作检测依据，具体的动作识别方法在第 5 节；第 6 节展示 IXMAS 数据集上的实验结果；第 7 节展示在运用深度数据的实时系统上实验的效果；第 8 节在最后总结全文并提出今后的研究目标.

2　研究现状

专注于动作特征值提取与类型识别的方法通常

实施在已被预先分割的视频片段上,其假定了在动作识别之前,分割过程不需要对动作内容进行判断,是独立完成的.但是,并非所有的研究观点都支持动作分割与识别是两个独立的过程.所以,近年来,动作时域分割领域的主要方法从是否基于内容方面,可以分为两大类[4]:

一类是不分析运动内容的方法,将分割过程独立于动作识别过程,通过各帧的某种统计特征值变化的特殊点来确定动作分歧点.这种方法使用若干个可表示运动信息的统计量作为依据,分析其相对于时间的函数,通过寻找极值点、不连续点来分割人体运动.从计算效率和系统实现角度来讲,在识别动作类型之前就能有效分割动作是十分有利的.由于不关心运动具体内容,无需先验知识,标记分割位置的效率远远高于标记种类,所以这类方法实现较为简单,运行速度快.光流能量[2]是比较经典的运动特征值,Weinland 等人[7]将能量随时间变化的峰值作为分歧点,较好地分割出了简单动作,如抬腿、放腿、弯腰、伸手等单向动作碎片.但这类方法所统计的特征值,或基于动静态姿势切换所产生的信息变化,或基于肢体部位运动轨迹,在分析具有多个部位运动信息的复杂动作的分歧点时,概括的信息比较片面,无法实现高准确率的分割.

另一类基于运动内容的方法将识别和分割的过程结合起来,需要对待识别动作进行建模,通过分类不同运动内容来实现分割,这类方法根据建模和分类方式不同可分为以下几类:

(1)聚类分析方法旨在把视频各帧分组,整个视频序列的各帧根据相似度距离被分到不同组,每一组被表示为一个动作.这类方法基于视频的所有帧建立分析动作类别,具有很高的扩展性和鲁棒性,同时对先验知识的依赖较少.其中,Zelnik-Manor 等人[8]提出的针对行为内容的统计分析方法,并不局限于帧与帧之间的空间相关性,取得了较好的聚类效果.但这种从单帧的角度出发进行分析的方法,难以考虑到时域上的动态特征变化,此外,其需要将整个视频序列信息输入后进行识别,这限制了这种方法只能在静态的视频分析应用中发挥作用.

(2)基于模板匹配的方法需要预先建立动作模板,通过计算相似度来检测是否有动作发生.这类方法训练较为简单,识别运算的复杂度仅与相似度运算有关,实时性较好.但这类方法受到模板的制约,模板动作的区分度和代表性将对识别结果产生很大影响.Lv 等人[9]的"Action Net"用图形形式,从空间相关性角度优化组织了各类人体动作模板,使得各个模版之间具有良好的区分度,但总体而言,这类方法更适合处理简单动作,处理复杂动作时,其常常难以构建具有较大差异的模板,鲁棒性较差.

(3)基于机器学习的方法在处理复杂动作时,不仅表现出较高的准确率同时还有极高的稳定性和鲁棒性.由于需要训练分割分类器,这类方法需要大量的被标记的训练数据,未降维的高维数据将使学习和识别速率降低,而经过特征提取后的数据可在一定程度上提高运算实时性.其中 Hoai 等人[4]提出的空间词袋模型,提取运动内容的特征向量,从而提高了 SVM 分类器效率.

本文采用分歧点检测的方法抽取关键帧,再通过基于模板匹配的方法检测新出现的动作并进行识别,这主要是因为模板匹配方法在应用于人机交互的增量式视频中时,在实时性方面十分优秀.

3 关键帧提取

Lu 等人[5]提出了一种有效的在连续动作视频序列中提取关键帧的方法,这个方法来源于 Ho 等人[10]提出的在数据流中通过鞅框架来检测数据的可替换性方法,而鞅框架则是 Vovk 等人[11]首次提出.这种方法通过检测视频数据的鞅之间的可替换性,以可替换性的改变帧,即对前序序列不再有可替换性的帧作为关键帧.Ho 等人[10]认为,如果一个数据流之间的 n 个值$\{s_1,s_2,\cdots,s_n\}$是可替换的,那么则有相应可替换性的鞅值函数$\{M_1,M_2,\cdots,M_n\}$是一个鞅过程,即满足:

$$M_n=E(M_{n+1}\mid M_1,M_2,\cdots,M_n) \tag{1}$$

鞅过程的含义在于:根据目前所知的信息,某变量在未来的预期就是此时刻的值.如果基于未来实际观测值认为该过程仍然成立,那么可认为数据流$\{s_1,s_2,\cdots,s_n\}$各个数据差异性不大,数据分布符合表示同一状态的概率分布条件;反之,若出现观测值使得鞅过程不成立,则可认为变量的值已被约束于另一概率分布条件,即数据流的状态发生改变.Vovk 等人[11]提出此鞅过程成立的假设中,存在 Doob 不等式作为拒绝的条件:

$$P(\exists n\mid Mn\geq\lambda)\leq 1/\lambda \tag{2}$$

而其中 λ 的值可视作鞅过程成立的条件,即阈值 λ 值越大代表可接受的可替换性变化越大.本文中通过设置 λ 值来控制提取关键帧的数量,当鞅过程假设被拒绝,则当前数据被视为关键帧.

3.1 各帧的源数据提取

对于视频数据流 $F=\{f_1,f_2,\cdots,f_n\}$,对每一帧

进行背景去除和中心化后，提取人体姿势轮廓图 $D=\{d_1,d_2,\cdots,d_n\}$. 由于本文不涉及背景去除和人体追踪方法，所以这一部分的工作被考虑为前期工作，轮廓图序列 $D=\{d_1,d_2,\cdots,d_n\}$ 在本文中成为源数据输入.

3.2　各帧的奇异值计算

Lu 等人[5]认为视频序列信息的奇异值是数据流奇异值的一种特殊形式，代表了该帧相对于整体序列的特殊程度，包含相似姿势的帧则具有相近的奇异值.

Lu 等人[5]根据 Ho 等人[10]提出的针对整体数据流奇异值计算方法，提出了视频流奇异值的计算方法，但其方法针对的是整体视频，需要整体视频序列信息，并求每帧针对整体信息的绝对奇异值.

而在本文中，由于只有当前帧的前序序列是已知的，所以只能求局部信息的相对奇异值作为分歧检测的特征值，相对前序序列奇异值的计算方法如下：

相对于前帧序列 $D_{i-1}=\{d_1,d_2,\cdots,d_{i-1}\}$，新的数据帧 d_i 的奇异值 s_i 可计算为

$$s_i=s(D_{i-1},d_i)=\|d_i-u_{i-1}\| \tag{3}$$

其中 $\|\cdot\|$ 为欧式距离，μ_{i-1} 为前序序列 D_{i-1} 的中心均值，其可由以下方式计算：

$$\mu_{i-1}=\sum_{j}^{i-1}dj/(i-1) \tag{4}$$

但是，此均值公式在求局部解时需要多次遍历前序序列，增加了计算规模. 可将上式改写为

$$\mu_{i-1}=\frac{d_{i-1}+(i-2)\mu_{i-2}}{i-1},\ i=2,3,4,\cdots \tag{5}$$

其中 μ_{i-1} 为计算前一帧 d_{i-1} 的奇异值所计算的中心均值矩阵，为了提高计算效率，本文采用迭代方式计算均值，保存上一帧次均值计算结果用于下次计算，无需多次遍历前序序列.

3.3　鞅框架构建与判别关键帧

Vovk 等人[11]指出，在奇异值序列 $S=\{s_1,s_2,\cdots,s_n\}$ 的基础上，鞅值由一个概率值 $\hat{p}$ 构造. 该概率意义为当前数据中流至少能获得一个与实际观测值一样的值的概率. 该概率由下式构建：

$$\hat{p}_i(\{d_1,d_2,\cdots,d_i\},\theta)=\frac{\#\{j:s_j>s_i\}+\theta\#\{j:s_j=s_i\}}{i} \tag{6}$$

其中 $\theta\in[0,1]$，其在本文中设为常数 0.5.

由此可以构建对应的鞅 M_i 为 s_i 与 $\hat{p}$ 的函数，表示 s_i 在当前奇异值序列中的可替换性：

$$M_i=\prod_{j=1}^{i}(\varepsilon\hat{p}_j^{\varepsilon-1}) \tag{7}$$

其中 $\theta\in[0,1]$，$\varepsilon\in[0,1]$，其产生方法由不同类型的鞅框架决定，本文参照 Lu 等人[5]在提取视频关键帧所使用的鞅框架计算方法，分别设这两个值为常数 0.5 和 0.8，以简化算法.

由上式可知，鞅值 M_i 的值也可由迭代方式计算，无需重复遍历前序序列：

$$M_i=\varepsilon\hat{p}_i^{\varepsilon-1}M_{i-1} \tag{8}$$

以 Doob 不等式作为判断关键帧的依据，如果 $M_i<\lambda$，则鞅过程假设成立，否则当前帧被认为是关键帧，而前序序列被清空，关键帧检测过程重新开始. 显然，$\lambda>1$ 是必须的，否则所有的帧将被判断为关键帧.

由图 3 可知，λ 的取值直接影响到关键帧的数量，对于更低的阈值，更多的大鞅值被检测到，从而产生了更多的关键帧. 更多的关键帧意味着回溯过程的步数增加，从而增加了运算负担. 但是更多的关键帧也提供了更全面的信息，减少了后续动作表示中的信息丢失.

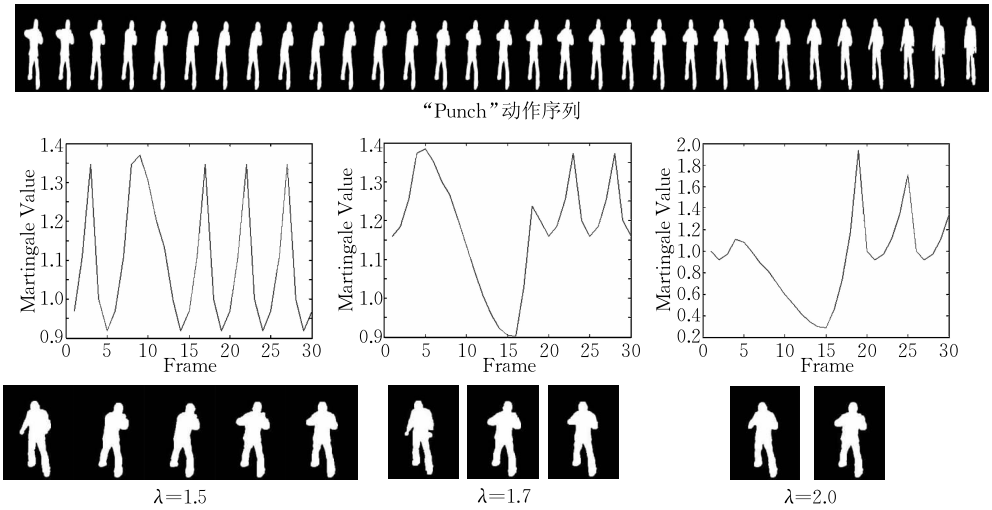

图 3　不同的 λ 值条件下的关键帧提取

4 动作表示与分割依据

4.1 运动历史图像

为了保存从一个关键帧到另一个关键帧之间动作的信息,包括动作移动的方式和移动的位置,本文采用动作历史图像(MHI)表示动作.

在基于轮廓和姿势的动作识别中,MHI 有其独特的优势,因为它能考虑到一个动作在时间和空间上的相关性.更为突出的是,MHI 的数据维度是固定的,它只与构成它的每一帧的维度相关,而不随着帧数的增加而增加.对于两个关键帧之间的序列 $D=\{d_1,d_2,\cdots,d_i\}$,MHI 可用一个简单的替代和衰减运算得到[6]

$$H_\tau(x,y,i)=\begin{cases}\tau, & h(x,y,i)\\ \max(0,H_\tau(x,y,i-1)-1), & \text{其他}\end{cases} \tag{9}$$

当沿着关键帧对已发生的动作信息进行回溯时,相邻的多个 MHI 需要被拼接成一个 MHI 图像.将某序列 $D_b=(d_1,d_2,\cdots,d_k)$ 与相邻的前序序列 $D_a=\{d_1,d_2,\cdots,d_i\}$ 的 MHI 进行拼接,计算如下:

$$H_\tau(x,y,i+k)=\begin{cases}H_\tau(x,y,k), & H_\tau(x,y,k)>0\\ H_\tau(x,y,i)-k, & \text{其他}\end{cases} \tag{10}$$

之后的包含判别只关注动作历史图像中的运动过程而不是运动姿势,本文将此动态过程图像称 PMHI,定义为

$$P_\tau(x,y,i)=\begin{cases}0, & H_\tau(x,y,i)=\tau\\ H_\tau(x,y,i), & \text{其他}\end{cases} \tag{11}$$

其中,动态过程有效像素数为运动能量 E:

$$E=\sharp\{j:P(x,y,j)>0\} \tag{12}$$

4.2 包含率

在回溯过程中,每当往前回溯一个关键帧,都要判别此时动作是否已完成.本文采用运动图像相对于模板的包含率来作为动作发现的初步判断依据.一个待判别动作 P_i 相对于一个模板动作 P_t 的包含率 $r(i,t)$ 定义为

$$IN=\sharp\{(x,y):P_i(x,y)>0\&P_t(x,y)>0\},$$
$$OUT=\sharp\{(x,y):P_i(x,y)>0\&P_t(x,y)<0\},$$
$$r(i,t)=(IN-OUT)/E_k \tag{13}$$

其中 IN 代表待判别动作包含于模板的部分,而 OUT 代表其超出模板的部分(图 3).

动作模板由不同训练样本的同一个动作求均值得来,每个完整动作相对于均值模板的包含率的最小值作为动作检测初步判别的阈值.

虽然每一次回溯都有可能发现当前已发生动作,但由于动作数量远小于关键帧的数量,所以大部分情况下,回溯中所产生的动作过程 MHI 都并不代表已完成的动作,而通过包含率判别是否有可能完成的动作,产生分割的参考位置,其运算规模是固定的,即对所有模板扫描一遍,并在一定条件下求和.

由图 4 可知,大幅度动作相对小幅度动作模板也可能产生超过阈值的包含率,所以回溯过程中很可能多次出现不同的超过阈值的 PMHI,所以需要进一步准确地确定此待判别动作是否与模板动作为同一类别,所以超过阈值的待判别动作将会进一步被识别.但相对于精确识别算法,采用包含率进行初判别的优势在于:每一次回溯,针对 PMHI 的识别只需要遍历模板一次.虽然初步识别结果并不唯一,且准确性也较低,但在时间上,其计算损耗远远低于每次回溯就进行精确动作识别.本文通过分析包含率寻找分割参考位置,尽可能地检测到动作,然后通过更为精确的模板匹配方法进一步确定动作类别并确定分割点.

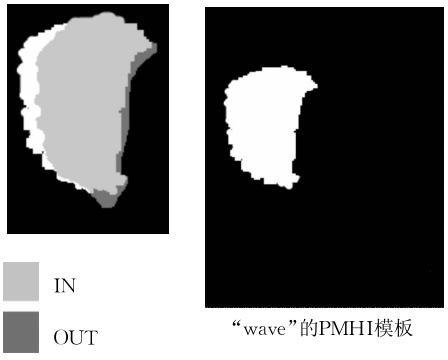

图 4 包含率计算

5 动作识别

本文采取基于 7 阶不变矩的特征值匹配方式,精确识别 MHI 所代表的动作类型.基于外形轮廓的动作表示方式,利用 7 阶不变矩[12]进行模板匹配来识别其种类是一种经典的方法.7 阶不变矩在对基于 MHI 的动作数据识别上,提取的特征值具有旋转不变性和尺度不变性.关于不变矩的定义和详细推倒过程见附录[12].

本文通过求待判别 MHI 与模板动作 MHI 的马氏距离来识别动作是否与模板动作为同一类.此马氏距离定义如下:

$$\gamma^2=(\boldsymbol{V}_i-V_t)^{\mathrm{T}}\boldsymbol{c}^{-1}(\boldsymbol{V}_i-V_t) \tag{14}$$

其中 γ 是马氏距离,$\boldsymbol{V}_i$ 是待识别动作 MHI 的不变矩

向量，V_t是模板动作 MHI 的不变矩均值，$\boldsymbol{c}$ 是模板 7 阶矩的协方差矩阵. 在识别前，各类模板的最优阈值由经典的 AdaBoost 算法得出. 如果待识别动作与模板的马氏距离在规定的阈值范围内，则可认为是一次成功的匹配，一个动作可以被最终识别并确定，如果有多个待识别动作与其模板的马氏距离在阈值范围内，那么拥有最小距离的那个动作为识别结果. 由于一个动作已经产生，将此动作分割后，待判别的数据序列将被清空，那么新的关键帧出现后，回溯过程不会再回溯到这个动作的序列中来.

6　IXMAS 数据集上的实验

6.1　数据集与实验环境

本文的方法在多镜头 IXMAS 数据集上进行了实验，包括其中的 132 个视频序列(33 次拍摄×4 个镜头). 每个序列中包括 12～14 个连续动作，其中的 13 个动作被选为识别对象. 在实验中，较为有代表性的 4 个人的动作被用来作为训练模板，其余作为测试对象.

运行实验程序的 PC 拥有 Intel Core i7 CPU 3.40 GHz，RAM 16 GB. 实验在 Matlab 平台上进行，没有采用任何额外的并行优化方式.

6.2　运算实时性评估

被测试的运算时间从第一帧人体轮廓数据输入开始，到最后一帧的识别过程结束. 由于本文方法是沿着关键帧对视频序列进行回溯，那么关键帧的数量对运算的实时性有着较大的影响. 而 λ 是影响关键帧数量的主要因素，所以在不同 λ 值的情况下，实验计算了程序对每个序列的运行时间的均值. 结果如图 5.

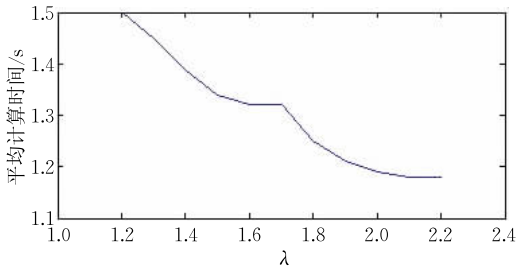

图 5　不同 λ 下的平均计算时间

从图 5 中可以看出，随着 λ 的增大，关键帧数目变小，而运算时间在缩短. 对于一段至少具有 1120 帧即至少 36 s 长的视频序列，其运行时间保持在 2 s 以下，这完全可以保证识别过程在 HCI 系统上运行时的实时性. 本文最终固定 λ 为 2，因为在反复的实验中，λ 处于这个取值层面时，关键帧数目和运行时间比较稳定，同时它也可保证比较适中的回溯跨度(即每一步跳跃的帧数)，从而有助于提高动作发现率.

6.3　动作发现与时域分割的效果

由于数据是增量式输入方式，发现动作只能根据当前已知的动作信息，而一旦分割后就不能再次回溯和修改，所以一旦发生有误差的分割，这些误差将很可能导致后续动作无法正确被发现，或者没有被发现. 在实验中，定义发现率为

$$\text{发现率}=\frac{\text{包括正确帧的动作数}}{\text{总动作数}} \tag{15}$$

被正确发现的动作并不一定被精确地分割，但却一定是在其发生的时刻被识别了. 对于人机交互的实际应用而言，重要的不是动作的信息是否完整，而是动作的内容，即类别或者含义，是否被及时的传达了. 所以发现率更能反映本文方法对不同类型动作的识别效果.

从表 1 中可以看出第一个动作 check watch 有着较高的发现率(93%)，因为它不会受到之前分割不准确的影响，而之后的动作如 cross arms 和 scratch head 的相似度比较高却又相邻，导致互相之间有较大影响从而导致发现率下降(低于 80%). 而之后的 turn around、walk 等等都有着较高的发现率(100%)，主要原因可能是其动作幅度较大，与其他动作或者什么都不做有着较为明显的区别. 实验的总体平均发现率约为 88%.

表 1　不同镜头下平均动作发现率

动作类型	镜头 0	镜头 1	镜头 2	镜头 3	平均
check watch	0.82	0.91	1	1	0.93
cross arms	0.82	0.91	0.55	0.55	0.7
scratch head	0.73	0.91	0.82	0.64	0.77
sit down	0.82	0.82	0.91	1	0.89
get up	1	0.82	0.82	1	0.91
turn around	1	1	1	1	1
walk	1	1	1	1	1
wave	1	1	0.73	0.73	0.86
punch	0.82	0.82	1	0.82	0.86
kick	1	1	1	1	1
point	1	0.91	0.91	0.82	0.91
pick up	0.82	0.91	1	0.82	0.89
throw	0.55	0.64	0.73	0.73	0.66
平均	0.87	0.9	0.88	0.85	0.88

为了关注每一个被发现的动作具体有多少帧是正确地被分割了，定义分割识别的准确率为

$$\text{准确率}=\frac{\text{被正确分割的帧数}}{\text{动作的总帧数}} \tag{16}$$

图 6 中每行代表此动作中被识别成各类动作的帧数百分比，以此代表每个动作在帧数层面的平均

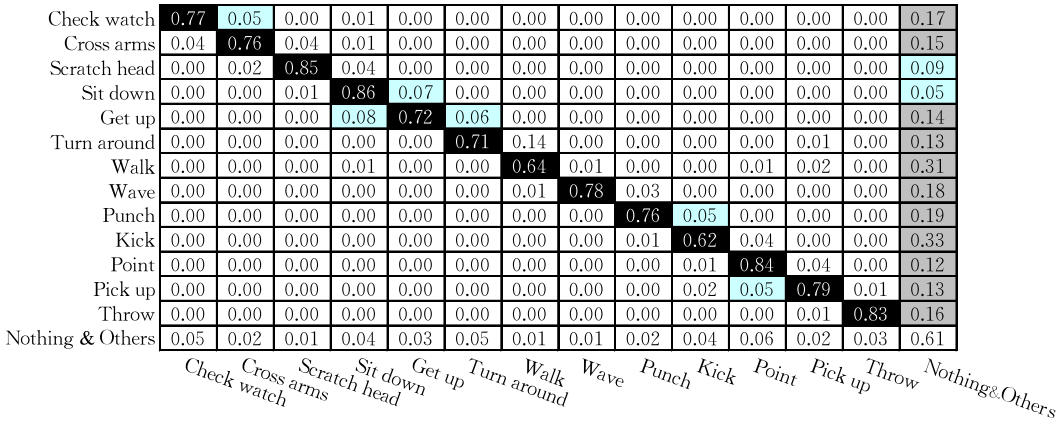

图 6 各个动作在帧层面的识别准确率与错误率

准确率和错误率.

从图 6 中可以看出,相邻动作之间的干扰是十分明显的,但是总体而言,在帧数层面上,识别率还是比较高的,尤其是 sit down、point 和 throw 这 3 个动作的准确率达到了 80%以上. Walk 与 Kick 的准确率最低(分别为 64%和 62%),这意味着它与相邻动作或者什么都不做的状态存在较大相关性,使得分割的位置更难以被准确找到. 其他幅度较大的动作也都能达到较高的准确率(高于 70%),这也意味着它们更容易被分割.

6.4 对 比

将本文的方法与其他 7 个在 IXMAS 数据集上的动作识别方法进行对比[5,7,9,13-16]. 表 2 展示了准确率对比. 这些算法的准确率来源于相应的文献中.

表 2 使用了 IXMAS 数据集的不同识别方法对比

方法	数据类型	镜头 0	镜头 1	镜头 2	镜头 3	镜头 4	平均
Weinland et al.[7]	已分割动作	—	—	—	—	—	93.3
Lv and Nevatia[9]	连续动作	81.5	82.1	80.1	81.3	78.4	80.6
Yan et al.[13]	已分割动作	72.0	53.0	68.0	63.0	—	64.0
Junejo et al.[14]	已分割动作	76.4	77.6	73.6	68.8	66.1	72.5
Liu and Shah[15]	已分割动作	76.7	73.3	72.0	73.0	—	73.8
Ramadan and Davis[16]	已分割动作	85.7	87.3	82.4	86.8	—	85.6
Lu et al.[5]	连续动作	77.6	79.3	78.8	83.1	83.6	80.5
本文方法	连续动作	75.5	77.1	74.9	76.1	—	75.9

从表 2 中可以看到,在平均准确率上,Weinland 等人在文献[7]中的效果最好,高达 93.3%,而本文方法的识别效果仅仅只能算中等水平(75.9%),但是大部分的方法都是针对于先分割好的动作序列. 同时包含分割和识别的方法[5,9],准确率达到 80%,但正如前文所述,这些方法需要一次性输入整个序列,并求整体的全局最优解,完整的获得每一个动作前后的信息有助于提高在帧层面的识别准确率,而 HCI 中的增量式视频数据处理只可能不间断寻找已知序列的局部最优解,无法获得后续未发生信息,所以这些方法难以应用到 HCI 系统中. 本文算法虽然准确率稍低,但是其不仅能针对连续动作序列进行分割识别,同时还能处理增量式数据,满足 HCI 中对动作分割识别的在线性要求的.

7 基于手势深度数据的在线实验

7.1 实验平台

实验在基于手势控制的交互原型系统上进行. 此系统应用 Kinect 作为动作信息输入设备,Kinect 可通过深度数据提供清晰的肢体轮廓. 深度数据的优势在于其不会受到光照因素的干扰,可通过深度层次的筛选,去除背景噪声,从而获得比普通彩色数据更为准确可靠的前景姿势信息.

实验所使用的手势交互系统是使用 C# 编写,其中拥有两个主要线程分别运行关键帧提取过程和动作分割过程. 这种并行结构使得关键帧提取过程不必等待分割过程的完成,以保证系统运行流畅.

由图 7 所示,系统通过提取近处深度信息,从而

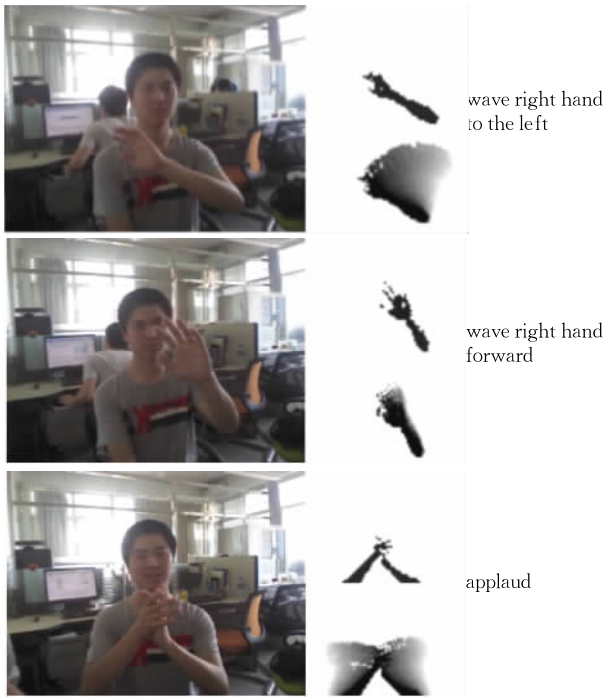

图 7　在线实验中的手势历史动作图像

只关注手部运动信息. 基于本文分割识别方法，当完整的动作被检测到之后，其类别判断结果将会被显示.

7.2　在线动作发现效果

实验系统对 10 个手部动作进行了建模，5 个用户以不同的顺序使用这些动作与计算机交互. 每一位用户进行 10 次实验. 动作的平均发现率见表 3.

表 3　在线实验中动作发现效果

动作类型	用户 1	用户 2	用户 3	用户 4	用户 5	平均
左手右挥	1.0	0.9	0.9	1.0	0.9	0.94
左手左挥	0.9	1.0	1.0	0.7	1.0	0.92
左手前挥	0.8	0.8	0.6	0.8	0.7	0.74
左手后挥	0.7	0.7	0.7	0.7	0.6	0.68
右手右挥	1.0	0.9	0.8	1.0	0.9	0.92
右手左挥	1.0	1.0	0.7	0.9	0.9	0.90
右手前挥	0.9	0.6	0.7	0.8	0.8	0.76
右手后挥	0.7	0.8	0.6	0.7	0.8	0.72
拍手	1.0	1.0	0.9	1.0	1.0	0.98
两手挥舞	1.0	0.9	0.8	0.9	1.0	0.92
平均	0.9	0.86	0.77	0.85	0.86	0.848

用户 1 是本文的作者，动作更为规范，所以实验结果相对较好(90%). 其他的用户在练习过后进行了实验，大部分的发现率都达到了 85%以上. 除此之外，向前挥手和向后挥手的发现率较低(低于 80%)，这主要是因为在深度层面上的改变信息没有被基于轮廓的 MHI 保留.

7.3　λ 对实时性和发现率的影响

由于 λ 值对关键帧数量有着决定性的影响，从而影响回溯过程的计算规模. 系统在不同的 λ 值情况下，进行了实时性测试，同时检测了 λ 值对发现率的影响.

表 4　不同 λ 情况下的实时性与发现率比较

λ	平均关键帧数量	平均每个动作的识别时间/ms	平均发现率
1.2	266.07	133.3	86.4
1.5	201.25	104.2	84.2
1.7	198.91	86.8	84.4
2.0	157.59	79.2	84.8
2.2	146.00	72.6	83.4

表 4 显示，λ 对关键帧数量有着明显的影响，随着 λ 的增大，关键帧数量减少，而相应的动作识别过程耗时降低，在 100 ms 以下的识别速度可以保证在 2 个关键帧产生的时间内，上一个动作已经被发现和识别，从而保证了在线交互的实时性. 同时，λ 的取值对发现率并无明显影响，这表明本文算法在此方面具有较高的稳定性.

8　结论与将来的工作

本文提出了一个面向增量式输入的连续动作序列的在线发现和时域分割方法. 对于一个不间断监视和观察人类行为的系统(如 HCI 系统)其输入数据是持续的、增量式的，该系统要求在一个动作发生后立即被发现并被分割. 首先，通过基于鞅框架的方法，每一个新输入的帧将被判断是否为关键帧. 然后，构建此关键帧与前一个关键帧之间的动作 MHI，并向前回溯拼接前序 MHI，通过计算对应各个模板的包含率，判断当前是否可能有动作符合分割要求. 最后通过计算可能动作与对于模板的 7 阶不变矩的马氏距离，判断其最终的动作类型.

本文方法中，关键帧的提取使得回溯过程无需逐帧进行，极大地提高了运算效率. 通过包含率进行初步分割，降低了每次回溯的运算规模，保证了实时性. 7 阶不变矩则为最终的分割和识别保证了较高的准确率. 其在 HCI 系统、安全报警系统和基于视频的事件记录系统上有良好的应用前景.

但该方法仍有以下问题：

(1) 实验表明本文方法在实时性上受到关键帧间隔的影响较大，关键帧间隔越大实时性越好. 但在识别效果方面，一定程度上关键帧间隔的增加对发现率的提高较小.

(2) 基于 MHI 的模板匹配算法受到模板和测试动作的约束程度高，建立模板的动作越多，代表性

越强,则测试效果更好,这也使得系统在构建稳定的动作模板时需要更多的资源,降低了可扩展性.

(3)另一方面,更为熟练的用户在测试中能获得较好的交互效果.所以本文方法推广性不足,鲁棒性低,对于区别度高的连续简单动作,该方法能获得更高的准确率,但对于相关性较大的复杂动作,该方法效果不佳.

将来的研究将关注识别算法上的优化,针对以上基于模板匹配方法的不足,拟通过基于机器学习的识别方法,在改进实时性的基础上,提高 HIC 系统整体的推广性和鲁棒性.

参 考 文 献

[1] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition//Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 2046-2053

[2] Fathi A, Mori G. Action recognition by learning mid-level motion features//Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

[3] Ahmad M, Lee S W. Human action recognition using shape and CLG-motion flow from multi-view image sequences. Pattern Recognition, 2008, 41(7): 2237-2252

[4] Hoai M, Lan Zhen-Zhong, De la Torre F. Joint segmentation and classification of human actions in video//Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3265-3272

[5] Lu Guo-Liang, Kodo M, Toyama J. Temporal segmentation and assignment of successive action in a long-term video. Pattern Recognition Letters, 2013, 34(15): 1936-1944

[6] Bobick A, Davis J. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-268

[7] Weinland D, Ronford R, Boyer E. Free viewpoint action recognition using motion history volumes. Compter Vision and Image Understanding, 2006, 104(2): 249-257

[8] Zelnik-Manor L, Irani M. Statistical analysis of dynamic actions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1530-1535

[9] Lv F, Nevatia R. Single view human action recognition using key pose matching and viterbi path searching//Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[10] Ho S. S, Wechsler H. A martingale framework for detecting changes in data stream by testing exchangeability. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(12): 2113-2127

[11] Vovk V, Nouretdinov I, Gammerman A. Testing exchangeability on-line//Proceedings of the 2003 International Conference on Machine Learning. Washington, USA, 2003: 768-775

[12] Hu M. Visual pattern recognition by moment invariants. IRE Transactions on Information Theory, 1962, 8(2): 179-187

[13] Yan P, Khan S M. Learning 4D action feature models for arbitrary view action recognition//Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-7

[14] Junejo I, Dexter E, Laptev L, Perez P. Cross-view action recognition from temporal self-similarities//Proceedings of the 2008 European Conference on Computer Vision. Marseille, France, 2008: 293-306

[15] Liu J, Shah M. Learning human actions via information maximization//Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8

[16] Ramadan S, Davis L. Action recognition using partial least squares and support vector machines//Proceedings of the 18th IEEE International Conference on Image Processing. Brussels, Belgium, 2011: 533-536

附 录.

对于一个 $M\times N$ 维的动作历史图像,其 $p+q$ 阶的矩 m_{pq} 定义为

$$m_{pq}=\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y)x^p y^q \tag{17}$$

其中 $p,q=0,1,2,\cdots$

其 $p+q$ 阶中心矩 μ_{pq} 定义为

$$\mu_{pq}\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y)(x-\bar{x})^p(y-\bar{y})^q \tag{18}$$

(x,y) 表示图像中的点,则物体中心点 $(\bar{x},\bar{y})$ 为

$$\bar{x}=\frac{m_{10}}{m_{00}},\ \bar{y}=\frac{m_{01}}{m_{00}} \tag{19}$$

其中:

$$m_{00}=\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y) \tag{20}$$

$$m_{10}=\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y)x \tag{21}$$

$$m_{01}=\sum_{x=1}^{N}\sum_{y=1}^{M}f(x,y)y \tag{22}$$

之后通过 0 阶中心矩 μ_{00} 对所有中心矩进行归一化处

理，可以得到规格化的各阶中心矩：

$$\eta_{pq}=\frac{\mu_{pq}}{\mu_{00}^{r}},\ r=\frac{p+q+2}{2},\ p+q=2,3,4,\cdots \tag{23}$$

Ho. M[13]通过 2 阶和三阶规格化中心矩的线性组合得到了 7 阶不变矩，其意义在于：图像的旋转、平移和缩放将不会改变这 7 个值：

$$v_1=\eta_{20}+\eta_{02} \tag{24}$$

$$v_2=(\eta_{20}-\eta_{02})^2+4\eta_{11}^2 \tag{25}$$

$$v_3=(\eta_{30}-3\eta_{12})^2+(3\eta_{21}-\eta_{03})^2 \tag{26}$$

$$v_4=(\eta_{30}+\eta_{12})^2+(\eta_{21}+\eta_{03})^2 \tag{27}$$

$$v_5=(\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]+(\eta_{03}-3\eta_{21})(\eta_{03}+\eta_{21})[(\eta_{03}+\eta_{21})^2-3(\eta_{12}+\eta_{30})^2] \tag{28}$$

$$v_6=(\eta_{20}-\eta_{02})[(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{30})^2]+4\eta_{11}(\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03}) \tag{29}$$

$$v_7=(3\eta_{21}-\eta_{03})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]-(3\eta_{12}-\eta_{30})(\eta_{03}+\eta_{21})[(\eta_{03}+\eta_{21})^2-3(\eta_{12}+\eta_{30})^2] \tag{30}$$

由于 7 阶不变矩的值太小，可以对其绝对值取对数，从而将其值的区分度体现出来：

$$v_k=\log|v_k|,\ k=1,2,3,4,5,6,7 \tag{31}$$

SHEN Qing, born in 1988, Ph. D. candidate. His main research interests include machine learning, human-computer interaction.

BAN Xiao-Juan, born in 1970, Ph. D., professor, Ph. D. supervisor. Her research interests include human-computer interaction, 3D visualization.

CHANG Zheng, born in 1986, Ph. D. candidate. His main research interests include machine vision and machine learning.

GUO Jing, born in 1990, M. S. candidate. His main research interests include machine vision and intelligence system.

Background

Temporal Segmentation of actions has been the topic of discussion in computer vision for a long time. Most Previous work of action analysis focused on action recognition and classification. During recent years, especially after the emergence a series of novel devices, the problems of the application in human-computer interaction has become more and more concerned. Temporal segmentation is one of those problems which restricts the understanding to human's behavior in HCI system. Currently, there was no standard method and theory to resolve or explain most problems in this area, although many approaches have been proposed, which performed with satisfied results in different experimental constraints. Most of experiments with action recognition technique were conducted under the well-controlled conditions, and the subjects of the study were the actions pre-segmented by human. However, without automatical segmentation, action recognition can not be widely used. Since HCI has been a popular application field of the computer vision, the more rigorous demands to the action analysis have attracted many attentions: a series of task should be complished immediately without the follow-up information, including action detection, segmentation, recognition and feedback. Most of recent HIC systems based on gesture, recognize the user's input through the position or angle of some part of body like hands or legs. These systems are unable to understand the complex actions.

Our work orients the particular requirements of segmentation in the continuous video of HCI system, proposes a optimization on characteristic frame selection and backtracking detection, which solves the problems of timeliness and realtimeness of on-line segmentation.

Our work belongs to the research project of "Multidimensional Virtual Sense Oriented Action Modeling and Behavior Understanding", which is supported by the National Natural Science Foundation of China (No. 61272357). The target of this project is to solve the difficulties listed above in multidimensional virtual sense system research and production. It can advance the experience of the system users, realize the practical natural human-computer interaction. Furthermore, it can provide the theoretical and technical support to the related industrial development.

Our group has achieved the phased objectives in 3D human movement modeling, action recognition invariant to background influence, depth data collection and so forth. Besides, action recognition based on the Extreme Learning Machine is also an important topic of our group in this research project.