基于特征点轨迹的动作识别

秦 磊[1] 胡 琼[1] 黄庆明[1),2)] 田 琦[3)]

[1)](中国科学院智能信息处理重点实验室 中国科学院计算技术研究所 北京 100190)

[2)](中国科学院大学 北京 100190)

[3)](美国德克萨斯大学圣安东尼奥分校 美国 78285)

摘 要 文中提出一种基于时空特征点轨迹的动作识别方法.首先为了克服局部时空特征时间信息缺失的问题,该方法采用 KLT 跟踪器对时空局部特征进行跟踪,将得到的时空特征跟踪轨迹作为基本的处理、描述单元.与局部时空特征相比,它能在更长的时间尺度上对运动进行描述,进而更好地捕获运动的动态变化与转变过程.其次在时空特征轨迹基础上,该方法提出了轨迹相对位置、相对速度关系元来对轨迹之间的关系进行建模.对轨迹之间的关系进行建模有助于捕获不同动作在特征分布上存在的一些比较稳定的模式.最后利用多核学习方法融合多种特征来训练动作分类器.在交互动作数据库上对提出的方法进行了实验,实验结果证明了方法的有效性.

关键词 计算机视觉;视觉特征提取;人体动作识别;特征点轨迹

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2014.01281

Action Recognition Using Trajectories of Spatio-Tempral Feature Points

QIN Lei[1)] HU Qiong[1)] HUANG Qing-Ming[1),2)] TIAN Qi[3)]

[1)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*University of Chinese Academy of Sciences, Beijing* 100190)

[3)](*University of Texas at San Antonio, TX* 78285, *USA*)

Abstract This paper proposes an approach to recognize human activities, which is based on tracking trajectories of local spatio-temporal feature points. To make up for the temporal information loss of local features, this paper uses the KLT feature tracker to track each spatial-temporal local feature and treats the tracked feature trajectory snippets as the basic processing and describing unit. Compared with local spatio-temporal feature, it can capture the motion information of an action pattern in a longer time scale and better describe the dynamic characteristics and transitions of motion. As to the relationship modeling among feature trajectory snippets, we believe that there exist some stable feature distribution patterns in an action video clip, which lie in the interconnection of position and velocity between local features, so we propose the relative position relation and relative velocity relation descriptors to capture this kind of relation. Experimental results on the UT-Interaction dataset are provided to demonstrate the effectiveness and robustness of our approach.

Keywords computer vision; visual feature extraction; human action recognition; trajectories of feature points

---

收稿日期:2012-08-27;最终修改稿收到日期:2014-01-16.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316400)、国家自然科学基金(61025011,61003165,61133003,61332016,61202234)、北京市自然科学基金(4111003)及高等学校博士学科点专项科研基金(20100142120029)资助.**秦 磊**,男,1977 年生,博士,副研究员,中国计算机学会(CCF)高级会员,主要研究方向为计算机视觉与模式识别.E-mail:qinlei@ict.ac.cn.**胡 琼**,女,1986 年生,博士研究生,主要研究方向为计算机视觉与模式识别.**黄庆明**,男,1965 年生,博士,教授,博士生导师,国家杰出青年科学基金获得者,主要研究领域为多媒体分析、图像处理、计算机视觉、模式识别等.**田 琦**,男,博士,教授,主要研究领域为多媒体分析、计算机视觉、模式识别等.

1 引 言

随着数码相机等视频获取设备的大众化及网络带宽的迅速增长，视频已经成为网络信息的主体成份，其中多数记录的是人的日常生活.因而对视频中的人体动作进行识别具有重要的意义[1].

在真实自然场景下，由于存在背景复杂、摄像机运动、遮挡和物体变化等因素，使得如何提取“好”的特征以及获取鲁棒的特征表达，对动作识别至关重要.时空局部特征和“视觉词袋”的表示方法在人体动作识别领域得到了广泛的应用.但是，这个框架通常完全依赖于单个时空局部特征的区分能力，忽略了时空局部特征的时空关系中可能蕴藏的有用信息，而局部特征能够表达的空间范围和时间范围都非常有限.

已经有一些研究工作通过对时空兴趣点邻域内的时空上下文信息进行建模试图构造相对鲁棒的视频描述子[2-3]，但是这些方法用到的时空上下文信息仍然有限.因此，在本文中我们试图找到一种能够在更长的时间尺度上对时空上下文进行建模的方法，即基于特征点跟踪轨迹以及轨迹之间关系描述的动作识别.

这种想法源于一个心理物理学实验，文献[4]中证实人能够通过人体关节点的运动轨迹来准确区分不同的人体动作，如图1所示，图1(a)～(h)表示不同时刻记录的人体上关节点的位置，图1(i)则通过浅色线来表示关节的局部运动信息，仅仅利用图1(i)中的信息人就可以准确的区分不同的人体动作.基于此，我们提出了一种基于特征点跟踪轨迹的人体动作识别方法，一方面克服了局部时空特征的时序信息缺失问题；另一方面有效利用了人能够通过关节点运动轨迹准确区分不同动作的视觉特性；同时也在一定程度上起到了噪声滤除的作用，因跟踪轨迹较短的特征点被滤掉了，而这些特征点往往是由于环境噪声或者其他因素造成的视觉上的突变.

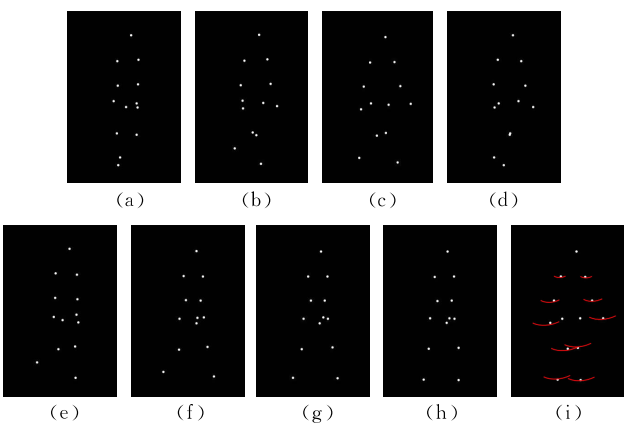

图1 人体关节运动的视觉感知

相关的研究工作有：文献[5]中提出对SIFT特征点进行跟踪，并利用马尔可夫链对轨迹进行建模，将最后得到的稳态分布作为固定维数的速度描述，但是这种稳态分布更像是一个速度的直方图而不是时间上更加结构化的视频描述.文献[6]将跟踪的关键点的速度用于产生式模型中进行动作识别.文献[7]提出了序列编码图(Sequence Code Maps)和相对位置概率，得到了两两特征之间关系简单高效的表达.这些工作都只是利用了轨迹的信息来进行动作识别，它们之间的差别在于对轨迹数据的不同描述方式.与上面的工作不同，本文中提出对轨迹与轨迹之间的关系进行建模以及基于此构建的轨迹关系元描述子，以增加在相对复杂动作(如两人交互动作)上识别的准确度.此外，轨迹特征和局部特征描述了动作的不同特性，两种特征是彼此互补、相互增强的.联合使用多种特征可以更全面地表示动作的特点，从而提高识别性能.实验结果也证明了这一点.

与国内外同类研究工作相比，本研究的贡献主要体现在以下几个方面：(1)利用了心理物理学实验的结论，增加了一般时空局部特征描述子缺失的时序信息；(2)提出了一种对特征点轨迹之间时空关系进行建模的方法，能够有效地编码相似轨迹之间的差异信息，并且该处理方法对其他基于轨迹的分析方法具有普适性；(3)在交互动作数据上进行了测试，并取得了很好的结果.

2 时空特征轨迹的获取

由引言部分的分析可知，为了描述一段动作视频的运动信息，需要对视频中提取到的关键点进行跟踪，该过程主要包括两步：第1步是决定跟踪哪些特征，第2步是跟踪过程本身.对于步骤1，可能考虑的特征点有视频帧上的角点或是SIFT特征点，或是视频中的时空兴趣点[8]、cuboid特征点[9]等.在动作识别中，我们更关注的是运动信息和运动模式，因此，本文选择后者，且Laptev提出的时空兴趣点相对cuboid特征比较稀疏，本文采用的是Dollar提出的cuboid特征.对于步骤2，常见的特征点跟踪方法包括：基于光流法的特征点跟踪方法、KLT特征点跟踪[10]以及更加复杂的基于卡尔曼滤波或是Camshift的特征点跟踪算法等.在文献[11]中，

Matikainen 等人利用 KLT 跟踪器就能够获得多个特征点的轨迹. 因此本文采用 KLT 跟踪算法.

在本文中，我们用 cuboid 时空特征检测器得到的特征点来初始化 KLT 跟踪器，对每一个特征点进行跟踪得到对应的轨迹，如图 2 所示. 在第 3 节中将介绍特征点轨迹的描述方式以及特征点轨迹之间关系的建模与描述.

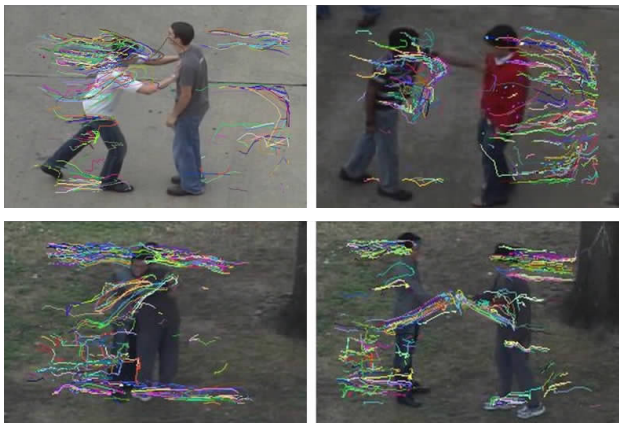

图 2　特征点跟踪轨迹

3　时空特征轨迹的描述

基于轨迹的动作识别在近年来得到了越来越广泛的研究[5,11-12]，算法的差异体现在如何编码一条轨迹的动态特性以及后继的一些处理上. 常用的轨迹的几何属性有：轨迹的速度和位置、轨迹的曲率和轨迹的分割等. 由于对不同的特征点跟踪到的轨迹的长度是变化的，为了实现轨迹之间相似度的度量和距离的计算，需要提前对轨迹进行预处理，在本文中，我们将轨迹划分为等长的轨迹片段. 为了利用轨迹的时空上下文信息，我们对任一特征点的轨迹片段，在不同特征点轨迹的轨迹片段中选取与其最接近的 K 条轨迹片段，并对之间的差异进行“编码”，以期在特征描述子中融入更多的时间、空间几何结构约束信息，以提高最终动作识别的准确度.

对于轨迹的划分与轨迹片段的描述，轨迹之间关系的建模和轨迹关系元的构建，下面将依次介绍.

3.1　特征点轨迹的划分与描述

一个特征点的跟踪轨迹是由该特征点不同时刻的位置向量构成的序列. 由于跟踪到的特征点的轨迹长度不一致，为了处理的方便，我们将一条轨迹有重叠或是无重叠地划分为等长的轨迹片段. 假定特征点 i 在 t 时刻的位置为 $\boldsymbol{s}_i^t=(x_i^t,y_i^t)$，则以 t 时刻为起点，长度为 L 的轨迹片段的速度可以表示为

$$\boldsymbol{v}=\{\boldsymbol{s}_i^{t+1}-\boldsymbol{s}_i^t,\boldsymbol{s}_i^{t+2}-\boldsymbol{s}_i^{t+1},\cdots,\boldsymbol{s}_i^{t+L}-\boldsymbol{s}_i^{t+L-1}\}\quad(1)$$

由式(1)得到的是一个长度为 $2L$ 的向量，是该轨迹片段上各个时刻 x 方向、y 方向速度的级联，此向量被称为轨迹元(trajecton)[11]. 从轨迹片段的划分和轨迹元的构造可知，对于一条长度大于 L 的特征点跟踪轨迹，其上任意长度为 L 的时间窗都可以形成一个轨迹片段，得到一个长度为 $2L$ 的特征描述向量. 在得到轨迹元的描述后，我们对轨迹元进行聚类和量化.

值得注意的是，在这种对轨迹的描述方式中，没有显式地考虑一个视频动作在时间上或是空间上的尺度变化，如果某特定动作以不同的尺度或是执行速度在视频中出现，则这些动作开始可能以不同轨迹元的集合进行表示，但是在分类阶段，我们通过给这些同类动作不同实例统一的标号来训练动作模型. 因此，如果训练样本包含的实例足够多，实例的变化足够丰富，在一定程度上可以容忍这种动作时空尺度的变化. 当然，如需更精细地处理这种尺度变化，可能需要用 DTW(Dynamic Time Warping)、HMM(Hidden Markov Model)等复杂的算法对时空数据进行预处理和对齐.

3.2　轨迹片段之间的关系建模

“上下文”这一术语被广泛地应用于计算机视觉与多媒体领域，但是由于其面向特定问题的属性，一直以来没有很明确的定义. 对于三维时空里动作识别这一问题，“上下文”被认为是任何包含一定的时空结构与变化、相对位置、全局或者半局部的统计特征(如灰度级、梯度和颜色等)等的时空信息[5]. 由于“上下文”能表达出运动本质的动态特性以及结构信息，对于无约束场景下的动作识别至关重要. 因此，我们提出一种基于轨迹“相似性”显式地对“相对位置”和“相对速度”进行编码的轨迹关系元特征.

3.2.1　轨迹片段间相似度的度量

从 3.1 节的介绍可以看出，一条特征点的跟踪轨迹可以表示成由一系列三元组$(\boldsymbol{s},\boldsymbol{v},t)$构成的序列，其中 $\boldsymbol{s}$ 表示跟踪到的特征点在 t 帧中的位置向量，$\boldsymbol{v}$ 表示特征点在 t 帧中的速度向量，这两个量都包含 x 和 y 两个方向的分量. 自然地，对两轨迹片段间相似度的度量既包括对他们位置上的邻近性的度量，也包括对他们在速度上的相似性的度量. 假定一特征点的轨迹片段 T_i 与另一特征点的轨迹片段 T_j 在时间上的重叠区间为 Γ，我们利用扩展的 McPhail 和 Wohlstein 距离来对特征点轨迹之间的距离进行度量[13]：

$$w_{ij}=\frac{\sum_t w_{ij}^t}{\rho_{ij}|\Gamma|},\quad i\neq j,\ t\in\Gamma \tag{2}$$

$$w_{ij}^t=\alpha N(\|\boldsymbol{s}_i^t-\boldsymbol{s}_j^t\|)+(1-\alpha)N(\|\boldsymbol{v}_i^t-\boldsymbol{v}_j^t\|) \tag{3}$$

$$\rho_{ij}=\sum_t \delta_t(i,j) \tag{4}$$

$$\delta_t(i,j)=\begin{cases}1, & \|\boldsymbol{s}_i^t-\boldsymbol{s}_j^t\|<H_s,\ \|\boldsymbol{v}_i^t-\boldsymbol{v}_j^t\|<H_v\\ 0, & \text{其他}\end{cases} \tag{5}$$

式(3)中 $N(\cdot)$是一个归一化操作,将数据线性地归一化到区间[0,1],并用权重系数 α 将空间上的邻近性和速度上的一致性融合到两轨迹片段 t 时刻的距离 w_{ij}^t 中. 式(5)中 H_s和 H_v是两个阈值. 对于两段轨迹片段,我们将它们在两者的重叠区间内的两两之间距离做平均,并将结果除以 ρ_{ij}(两轨迹片段上特征点 i 和特征点 j 之间的距离小于 H_s并且特征点 i 和特征点 j 之间的速度差小于 H_v的次数). 对于那些在空间上比较邻近,且在时间上以相似的速度一起运动较长时间的特征点,他们之间的距离 w_{ij} 比较小. 我们期待通过这种方式找到动作视频中比较稳定的运动模式,并且这种相似度度量对跟踪误差比较鲁棒. 尽管在特征点跟踪的过程中自动获取的特征点轨迹可能包含噪声,但是通过考虑较长时间上的一致性,我们仍然能够获得动作视频中比较稳定的运动模式. 如图 3 所示,其中长的直线或是

图 3 稳定动作模式示例(图中右列给出了全部特征点的跟踪轨迹. 左列是考虑轨迹在较长时间上的一致性,获得的动作视频中比较稳定的运动模式)

圆弧轨迹描述了特定的运动模式(推人、踢腿、握手).

3.2.2 轨迹关系元的构建

通过 3.2.1 节介绍的轨迹片段之间相似度的度量准则,我们可以找到与一条轨迹片段最接近的 K 个近邻,本文实验室中 K 取 3,这种最近邻不仅是一种时空位置意义上的邻近,而且是倾向于选取动作识别中比较稳定一致的运动模式.

前面提到一条轨迹自身的描述能力比较局限,因此我们在此构建一种轨迹上下文的建模方式,这种建模方法的主要思想是对轨迹与其 K 近邻之间的"残差"进行编码,得到的是一种对轨迹之间关系的描述. 比如,一种动作模式形成的特征点的运动轨迹比较一致,则得到的残差项较小,而当一种动作自身包含很多不一致的细节运动时,得到的残差项较大,这种方式是对某类动作中运动模式的一种精细编码,是基本的轨迹特征的一种补充. 如表 1 和表 2 所示,分别是对轨迹之间相对位置和相对速度的编码.

表 1 轨迹片段之间相对位置关系

T_0	$\boldsymbol{s}_0^1$	$\boldsymbol{s}_0^2$	…	$\boldsymbol{s}_0^L$
T_1	$N(\|\boldsymbol{s}_1^1-\boldsymbol{s}_0^1\|)$	$N(\|\boldsymbol{s}_1^2-\boldsymbol{s}_0^2\|)$	…	$N(\|\boldsymbol{s}_1^L-\boldsymbol{s}_0^L\|)$
T_2	$N(\|\boldsymbol{s}_2^1-\boldsymbol{s}_0^1\|)$	$N(\|\boldsymbol{s}_2^2-\boldsymbol{s}_0^2\|)$	…	$N(\|\boldsymbol{s}_2^L-\boldsymbol{s}_0^L\|)$
…	…	…	…	…
T_K	$N(\|\boldsymbol{s}_K^1-\boldsymbol{s}_0^1\|)$	$N(\|\boldsymbol{s}_K^2-\boldsymbol{s}_0^2\|)$	…	$N(\|\boldsymbol{s}_K^L-\boldsymbol{s}_0^L\|)$

表 2 轨迹片段之间相对速度关系

T_0	$\boldsymbol{v}_0^1$	$\boldsymbol{v}_0^2$	…	$\boldsymbol{v}_0^L$
T_1	$N(\|\boldsymbol{v}_1^1-\boldsymbol{v}_0^1\|)$	$N(\|\boldsymbol{v}_1^2-\boldsymbol{v}_0^2\|)$	…	$N(\|\boldsymbol{v}_1^L-\boldsymbol{v}_0^L\|)$
T_2	$N(\|\boldsymbol{v}_2^1-\boldsymbol{v}_0^2\|)$	$N(\|\boldsymbol{v}_2^2-\boldsymbol{v}_0^2\|)$	…	$N(\|\boldsymbol{v}_2^L-\boldsymbol{v}_0^L\|)$
…	…	…	…	…
T_K	$N(\|\boldsymbol{v}_K^1-\boldsymbol{v}_0^1\|)$	$N(\|\boldsymbol{v}_K^1-\boldsymbol{v}_0^1\|)$	…	$N(\|\boldsymbol{v}_K^L-\boldsymbol{v}_0^L\|)$

表 1 与表 2 中,T_0 表示当前考虑的轨迹片段,$T_i(i=1,2,\cdots,K)$表示 T_0 的第 i 条最近邻轨迹片段,$\boldsymbol{s}_i^t$与 $\boldsymbol{v}_i^t(i=0,1,\cdots,K;t=1,2,\cdots,L)$分别表示轨迹片段 T_i上时刻 t 的位置向量与速度向量,$N(\cdot)$是一个归一化操作. 表 1 中各元素为归一化的位置偏差,而表 2 中各元素为归一化的速度偏差. 我们将表 1 与表 2 按行向量化为一个长度为 $K\times L$ 的一维向量,分别称为该轨迹片段的相对位置关系元(disRelation)与相对速度关系元(velRelation),这两种特征合称为轨迹关系元特征.

同样的,我们可以对轨迹关系元用 BoW(Bag of Words)模型,得到一段动作视频中相对位置关系元与相对速度关系元的统计规律,并方便与前面的时空特征、轨迹特征等放在同一处理流程下.

4　动作识别方法

基于前面提出的时空特征轨迹特征，我们提出了一种多特征融合的动作识别方法. 我们所利用的特征包括：(1) 对视频中时空局部立方体进行描述的 cuboid 特征；(2) 对这些特征点跟踪轨迹进行描述的轨迹元，它可视作这些特征点在时间上的延拓，能够有效地捕捉运动的动态变化与转变，是对动作模式的一种粗编码；(3) 对不同特征点轨迹片段之间时空关系进行刻画的轨迹关系元特征，包含轨迹相对位置关系元与轨迹相对速度关系元，这种特征可看作是轨迹元特征基础上对动作运动模式的一种精细编码. 这些特征从不同的侧面对动作视频进行描述，因此可以期望这些不同属性的特征在人体动作识别，尤其是对包含人与人交互的动作识别中起到互补作用.

由于在不同的视频片段中得到的上述 3 种特征的数目不同，为此我们引入了 BoW 方法[14]. 对每种类型原始特征分别进行聚类，得到能代表类别的字典，然后将原始特征量化并生成相应的统计直方图，这样每个样本就可以表示成不同特征类型的直方图向量. 最后把这些特征直方图通过多核学习的方式进行融合并用学习支持向量机分类器进行分类识别. 图 4 显示了我们提出的基于特征点跟踪轨迹描述的动作识别方法的框架图.

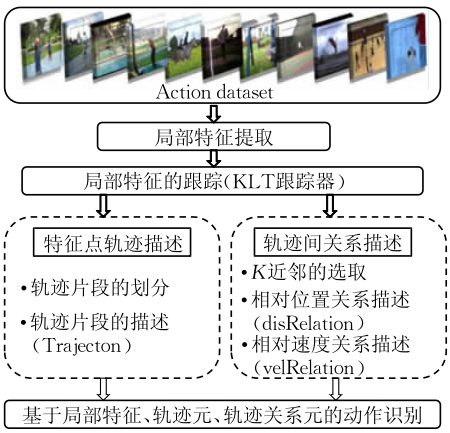

图 4　基于特征点跟踪轨迹描述的人体动作识别算法框图

5　实验与性能比较

本节将依次按照实验数据库、实验设计以及实验结果与分析对基于特征点跟踪轨迹描述的人体动作识别算法的性能进行介绍与分析.

5.1　实验数据库

在本文的实验中，我们采用文献[15]中提供的 UT-Interaction 数据库，见图 5 示意，这个数据库包含 6 类人与人之间的交互动作：Hand Shaking、Hugging、Kicking、Pointing、Punching 以及 Pushing. 整个数据库被分为两个子集：*Set*1(S1)与*Set*2(S2)，其中 S1 的背景是停车场，S1 中的视频是在较小的变焦率下拍摄的，背景几乎是完全静止的，只有较小的摄像机抖动；而 S2 是在一个有风的草地上拍摄的，背景本身存在一定的运动(如树的运动)，并且在拍摄的过程中包含更多的摄像机抖动. 两个场景有着完全不同的背景、尺度以及光照条件.

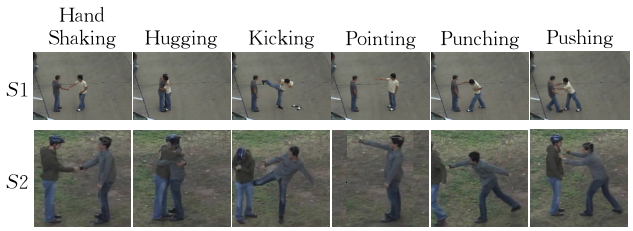

图 5　UT-Interaction 数据库

S1 与 S2 均包含 10 个不同的场景，每个场景包含几个人与人之间的交互动作. 在数据子集 S1 或是 S2 中，上述 6 类人与人之间的交互动作出现的次数分别为 10 次，各包含 60 段动作视频.

5.2　实验设计

若两段动作视频的 BoW 描述子为 H_i, H_j，它们之间的距离采用直方图交进行计算，公式如下：

$$K_{\text{intersect}}(H_i, H_j) = \sum_{b=1}^{k} \min(H_i(b), H_j(b)) \quad (6)$$

其中 b 是对直方图中 *bin* 的索引，k 表示该特征分布直方图中 *bin* 的总数.

对于本文前面介绍的时空局部特征、轨迹片段描述特征、轨迹片段之间关系描述特征，在融合方式上，采用多核 SVM 进行融合[16]：

$$K_{\text{Minkowski}}(i,j) = \sqrt[p]{K_{f_1}(H_i, H_j) \times K_{f_2}(H_i, H_j) \times \cdots \times K_{f_p}(H_i, H_j)} \quad (7)$$

式(7)中，$K_{f_i}(H_i, H_j)(i=1,2,\cdots,p)$表示的是基于一种特征的单通道核函数. 本文选用的是直方图交核函数[17]. 在本实验中，对每一个子集都采用 10 折的留一交叉验证进行测试.

5.3　实验结果与比较分析

根据上面的实验设计，本节分别对基于时空局部特征(cubiod 特征)、轨迹元、轨迹相对位置关系

元以及轨迹相对速度关系元等单种特征的人体动作识别以及它们之间不同组合的情况进行实验.

表 3 中给出了平均识别准确率,其中既包括每种特征单独使用的情况,也包括几种特征的组合结果,其中 f_1 表示 cuboid 特征; f_2 表示 trajecton 特征; f_3 表示 disRelation 特征; f_4 表示 velRelation 特征; f_i+f_j 表示不同特征的组合, i,j 属于 1～4,且 i 不等于 j. 表 3 中只给出了特征组合结果精度最好的前 4 种组合方式.

表 3 不同特征及其组合的动作识别结果

*Set*1		*Set*2	
Feature	Ave Accuracy/%	Feature	Ave Accuracy/%
f_1	78.33	f_1	73.33
f_2	83.33	f_2	81.67
f_3	56.67	f_3	56.67
f_4	50.00	f_4	48.33
f_1+f_2	88.33	$f_1+f_2+f_3+f_4$	83.33
$f_1+f_2+f_3$	86.67	$f_1+f_2+f_4$	81.67
$f_1+f_2+f_4$	83.33	f_2+f_3	81.67
f_1+f_3	80.00	f_1+f_3	78.33

从表 3 中可以看出,在单独使用各种特征的情况下,对单条轨迹进行描述的特征效果最好,时空局部 cuboid 特征次之,而对轨迹之间关系进行编码的特征再次之. 直观上看,这种对特征点进行跟踪得到的轨迹特征能够很好的捕捉到人体动作中的运动特征以及这种局部运动在时间上的动态演化,而时空局部特征只是对动作视频中局部运动的描述,没有考虑到时间上的动态性,因而前者的效果要优于后者. 至于轨迹相对位置关系元以及轨迹相对速度关系元特征,它们是不同轨迹片段之间关系的编码,是对轨迹以及局部运动特征的一种有效补充,在与上述两种特征进行融合的情况下,在不同数据子集上的实验性能都得到了不同程度的提高. 从表中我们可以看到,在两个数据子集上,通过特征融合得到的识别平均准确率比纯粹使用时空局部特征的情况下都提高了近 10 个百分点. 我们将这两个数据子集上得到的最好识别结果的混淆矩阵列于图 6 中,从混淆矩阵上可以看出,不同动作类别之间识别结果比较均匀,均在 0.80～1.0 之间,这种结果与数据的分布有关,因 UT-Interaction 数据库 *Set*1 与 *Set*2 中包含不同动作的数目是均衡的,6 类动作各 10 段视频,故出现图 6 中给出的实验结果也非常合理.

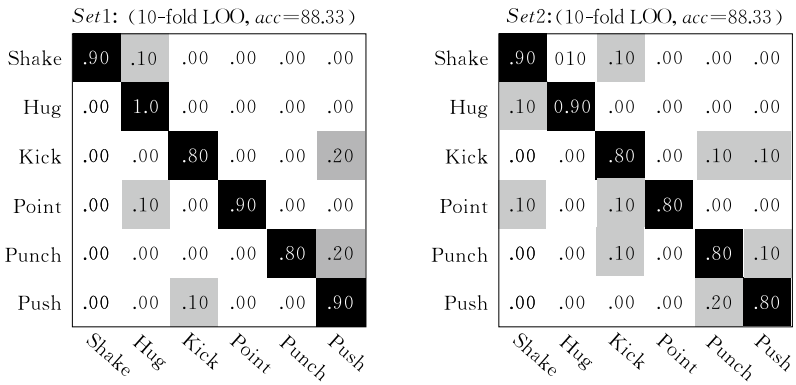

图 6 *Set*1 和 *Set*2 的混淆矩阵

为了更清楚地看到本文提出的特征对动作识别性能的提升,我们将本实验中得到的最好结果、仅采用时空局部特征的结果、采用文献[3]中方法的结果以及 SDHA2010 比赛中给出的最好结果(Team BIWI①)进行了比较与分析,详见图 7 与图 8.

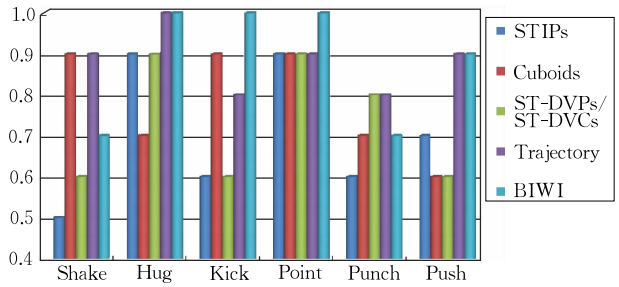

图 7 UT-Interaction *Set*1 实验结果比较

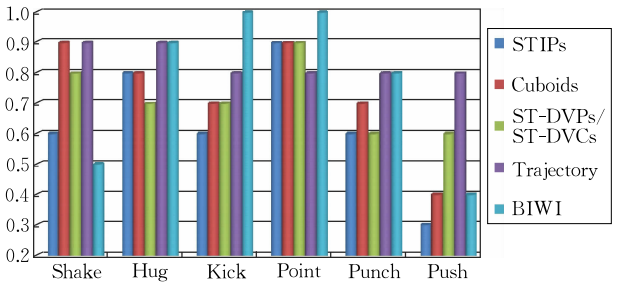

图 8 UT-Interaction *Set*2 实验结果比较

在此我们将采用时空局部特征的结果记为“STIPs”和“Cuboids”,它们分别对应于文献[8-9]提出的时空局部特征检测器检测的特征;将文献[3]

① http://cvrc.ece.utexas.edu/SDHA2010/

提出的特征记为“ST-DVPs/ST-DVCs”；将本文中实验得到的最好结果称为“trajectory”；“BIWI”代表SDHA2010 比赛中在该数据上得到的最好结果.

在 *Set*1 上，采用“STIPs”和“Cuboids”特征的平均识别准确率为 70.00%和 78.33%；在 *Set*2 上，对应于两种时空特征的平均识别准确率分别为 63.33%和 73.33%. 由此可见，对本数据而言，“Cuboids”特征要明显优于“STIPs”特征，这与 STIPs 比较稀疏有关.

对于同以“Cuboids”特征作为基本特征的方法，在 *Set*1 上，单用“Cuboids”特征的平均识别准确率为 78.33%，而本实验得到的平均识别准确率为 88.33%，与 SDHA2010 比赛中得到的最好结果 88.33%一样，准确率提高了 10 个百分点. 在 *Set*2 上，单用“Cuboids”特征的平均识别准确率仅 73.33%，BIWI 的识别准确率也只有 76.67%，而本实验方法达到了 83.33%，比 BIWI 以及基本特征分别提高了近 7 个和 10 个百分点. 从这些数据可以看到，首先本文中提出和构造的特征描述子在动作识别中是比较有效的，能够不同程度的对最终的动作识别准确率有所提升；其次与 *Set*1 相比，*Set*2 上的性能提升更为显著，这与数据本身的特点有关. 从前面对实验数据库的介绍可知，*Set*1 中主要是在静态背景下拍摄的含较小摄像机抖动的动作视频，而 *Set*2 是在背景本身包含运动的草地上拍摄的，且包含较剧烈的摄像机抖动. 由此可见，本文中构造的特征描述子对真实自然场景和各种噪声的干扰更加鲁棒有效.

仔细观察图 7 与图 8 中每一类动作在不同方法下的识别准确率，我们可以发现本文中提出的方法在 Shake、Hug、Punch 以及 Push，尤其是后面两种动作上识别准确率显著高于其他方法，而在 Kick 和 Point 这两类动作上，效果反而不如 BIWI. 通过对实验数据的分析发现，前 4 类动作主要是人与人之间的交互，而后 2 类动作虽然也涉及到两个人，但是动作的主要执行者却只有一个，比如，对于 Pointing 动作，从图 5 中可以清楚地看到，主要是动作的执行方将手指指向另一方，而另一方在此动作中几乎没有运动. 从本文前面介绍的特征构造过程可知，轨迹关系元主要是对不同特征点轨迹片段之间残差的一种编码，这种构造方式能够较好的捕捉两个或是多个运动对象之间交互运动所形成的一种稳定的共生模式，因此，本文提出的识别算法对前 4 类动作的识别效果有显著的改进，而在剩下的 2 类动作上改进效果稍差.

6 总　结

本文首先通过对时空特征点进行跟踪得到时空特征点的轨迹. 然后，将这一特征点的跟踪轨迹划分为等长的轨迹片段作为处理的基本单元. 这种将轨迹片段作为处理基元，而不是直接对时空局部特征进行处理的方式，可以有效地克服时空局部特征时间信息缺失的问题，为动作识别提供非常有用的线索. 对于轨迹之间关系的建模，主要是考虑到不同的动作类型，在提取到的特征点的分布上会出现一些比较稳定的模式，表现在特征点的位置和速度等之间存在一定关系，因而提出相对位置和相对速度两种轨迹关系元以捕获轨迹之间的这种时空关系. 本文提出方法的有效性在开源数据库 UT-Interaction 上得到验证，加入这些特征之后人体动作识别的平均准确率显著提高.

参　考　文　献

[1] Turaga P, Chellappa R, Subrahmanian V S, Udrea O. Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(11): 1473-1488

[2] Wu Xin-Xiao, Xu Dong, Duan Li-Xin, Luo Jie-Bo. Action recognition using context and appearance distribution features//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 489-496

[3] Hu Qiong, Qin Lei, Huang Qing-Ming, et al. Action recognition using spatial-temporal context//Proceedings of the International Conference on Pattern Recognition. Istanbul, Turkey, 2010: 1521-1524

[4] Johansson G. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 1973, 14: 201-211

[5] Sun Ju, Wu Xiao, Yan Shui-Cheng, et al. Hierarchical spatio-temporal context modeling for action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Florida, USA, 2009: 2004-2011

[6] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 104-111

[7] Matikainen P, Hebert M, Sukthankar R. Representing pairwise spatial and temporal relations for action recognition//

Proceedings of the European Conference on Computer Vision. Crete, Greece, 2010: 508-521

[8] Laptev I. On space-time interest points. International Journal of Computer Vision, 2005, 64(2): 107-123

[9] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features//Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing, China, 2005: 65-72

[10] Shi Jian-Bo, Tomasi C. Good features to track//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 1994: 593-600

[11] Matikainen P, Hebert M, Sukthankar R. Trajectons: Action recognition through the motion analysis of tracked features//Proceedings of the Workshop on Video-Oriented Object and Event Classification. Kyoto, Japan, 2009: 514-521

[12] Wu Xiao, Ngo Chong-Wah, Li Jin-Tao, Zhang Yong-Dong. Localizing volumetric motion for action recognition in realistic videos//Proceedings of the ACM Multimedia. Beijing, China, 2009: 505-508

[13] Ge Wei-Na, Collins R T, Ruback B. Vision-based analysis of small groups in pedestrian crowds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(5): 1003-1016

[14] Niebles J C, Wang H, Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision, 2008, 79: 299-318

[15] Ryoo M S, Aggarwal J K. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1593-1600

[16] Gehler P, Nowozin S. On feature combination for multiclass object classification//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 221-228

[17] Maji S, Berg A C, Malik J. Classification using intersection kernel support vector machines is efficient//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

QIN Lei, born in 1977, Ph. D., associate professor. His research interests include computer vision, and pattern recognition.

HU Qiong, born in 1986, Ph. D. candidate. Her research interests include computer vision, pattern recognition.

HUANG Qing-Ming, born in 1965, Ph. D., professor, Ph. D. supervisor. His research areas include multimedia video analysis, image processing, computer vision and pattern recognition.

TIAN Qi, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia analysis, computer vision and pattern recognition.

Background

Recognizing human actions is one of the hot spot issues of computer vision, which receives increasing attention due to its wide range application. Due to background clutter, occlusion, camera motion and illumination condition changes, there still remain some open issues in action recognition, and one of the most important issues is action representation. Many methods have been proposed in the last decades. The methods can be divided into four categories. First kind is static features such as shape, texture, and silhouette. Second kind is local features such as STIP, Cuboid. Third kind is the optical flow. Fourth kind is the middle level features such as various descriptions.

In this paper, we propose to use trajectories of local spatio-temporal feature points to recognize human activities. Compared with local spatio-temporal feature, it can capture the motion information of an action pattern in a longer time scale and better describe the dynamic characteristics and transitions of motion.

The research interests of our group are object detection, tracking and activity analysis. In the ten tasks of the Trecvid 2009 event detection competition, our group won four tasks. In the eight tasks of the Trecvid 2010 event detection competition, our group won two tasks. We use adaptive background modeling, body and head-shoulder detection, adaboost-based tracking, ensemble of one-vs.-all SVM and automata-based classifiers, effective event merging and post-processing.

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61003165, 61133003, 61035001, 61332016 and 61202234, and in part by Beijing Natural Science Foundation: 4111003.