doi:10.3772/j.issn.1002-0470.2020.02.009

基于 Kinect 和 ROS 的骨骼轨迹人体姿态识别研究①

胡敦利② 柯浩然③ 张 维

(北方工业大学现场总线及自动化北京市重点实验室 北京 100144)

摘 要 为了解决不完整人体动作识别的问题,在机器人操作系统(robot operating system,ROS)下提出利用骨骼节点向量的角度累计变化作为特征向量,并采用自适应能量的方法划分视频人体动作。在人体解剖学的基础上建立投影坐标平面,进行空间位置和骨骼角度的计算。通过时间金字塔方法对不同时间间隔的骨骼角度数据编码,形成多级特征向量更好地表示人体动作。在人体受遮挡情况下,使用扩展卡尔曼滤波预测骨骼节点坐标,提高骨骼坐标的准确性。该方法具有旋转、平移不变性,识别4种不完整人体动作的正确率达到了92.25%。

关键词 自适应能量;时间金字塔;扩展卡尔曼滤波;机器人操作系统(ROS);预测

0 引 言

机器视觉广泛应用于智能监控、家居安全、医疗监护、智能机器人以及运动员辅助训练[1]等领域。但机器视觉的应用不应该只限于简单的视频监控,而应该以视频中的内容为切入点,分析其中的数据,为人们提供更加智能化的服务,例如通过捕捉人的表情、动作等来预测人的行为或者意图。

经过多年的发展,人体姿势识别的方法大体分为模板匹配法[2]和状态空间法[3]。另一方面,在过去几十年的机器视觉和机器智能领域中,通常选取空间特性或者时间特性作为表征人体特征的描述性信息,再对其进行编码分析。在第1代 Kinect 发布后,微软剑桥研究院的 Shotton 等人[3]发表了关于利用骨骼信息进行姿势识别方法的论文,由此引领了一大批学者开始研究基于骨骼信息的人体行为识别。2014年,微软发布了第2代 Kinect,在原有的基础上提高了其性能,由此可以为人体姿态识别提供更加丰富和清晰的数据源。但在 Linux 系统和机器人操作系统(robot operating system,ROS)系统下,由于微软对于 Linux 系统下开发的支持并不好,所以需要使用第三方软件和一些中间件对 Kinect V2 进行开发。通过在 Linux 和 ROS 系统下开发,可以方便地在机器人上使用 Kinect V2 获取人体骨架信息,为之后的人体姿态识别提供坚实基础。目前已经有了很多人体动作识别的方法[4],它们大多数只对完整的动作做分类识别[5-7]。在日常生活中,摄像机捕捉到的人体动作往往是片段的,而非一个完整的行为动作,这就给人体动作识别带来了很大困难,片段的动作可能代表了与完整动作完全不同的含义。这就需要采用一种有效的特征提取方式。针对这个问题,本文设计了一种特殊的特征选取方法,利用人体的骨骼轨迹进行分析,最终形成有效的特征信息。

1 先期工作

1.1 获取骨骼信息

在 Linux 中使用 Kinect V2 需要使用第三方软件,libfreenect2 支持 RGB 图像、IR 和深度图像的获

① 国家自然科学基金(61573024)资助项目。

② 女,1967年生,博士生,副教授;研究方向:现场总线技术及智能控制;E-mail:hdl@ncut.edu.cn

③ 通信作者,E-mail:18610864910@163.com

(收稿日期:2019-03-19)

取。而在 ROS 中使用 Kinect V2 需要 iai _ kinect,它提供相机标定工具、深度数据配准工具,最重要的是它可作为 libfreenect2 和 ROS 的桥接工具。图 1 中给出了 4 种不同设备获取的人体骨架模型,OpenNI 画出了 15 个骨骼节点。Kinect V1 SDK 画出了 20 个骨骼节点,增加了对于手指、脚趾和髋关节中心的描述。Kinect V2 SDK 在一代的基础上又增加了 5 个关节点。动作捕捉系统(motion capture system,MoCap)则可以选择不同数量的骨骼节点。虽然 MoCap 系统可以提供更加丰富的骨骼节点,对于人体的描述更加精确,但它不可以用于移动的机器人平台上,基于这个原因本文选用了 OpenNI。

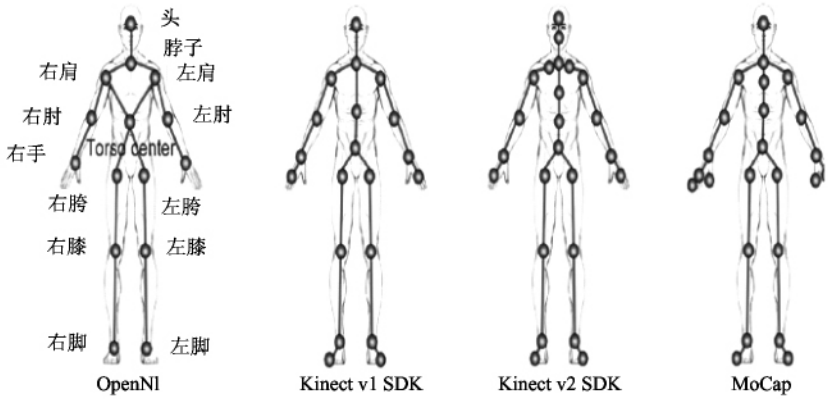

图 1 人体骨骼节点

动作特征指的是从人体动作序列中提取的可以正确描述人体动作状态的参数[8]。视频图像中的颜色、纹理、时间和空间等信息都可以作为人体姿态特征参数。通常人体姿势识别选用位置坐标、速度、相对角度等。文献[9]提出了一种基于人体骨骼关节点坐标的特征表示方法来进行动作识别。文献[10]提出了基于词袋模型的可以捕捉局部时空信息的特征表示方法,一方面词袋法不能准确地处理视角变化的问题,并且词袋法的计算量不适合实时处理的应用。另一方面之前基于词袋法的特征表示没有利用 Kinect V2 提供的深度信息。

不同于之前提到的基于骨架的特征表示方法,本文的方法不仅能够识别姿态还具有预测的功能。正是因为这种预测的能力,可以对正在进行的动作做出正确的分类,而不用等到整个动作做完。能够对未完成的动作进行预测的关键在于对时间信息的处理,以往的方法通常会采取在线的方式对机器学习方法进行扩展。与以往方法不同的是本文直接将时间信息融合到特征中,使得机器学习方法可以直接使用。

2 骨骼特征表示

2.1 生物解剖学基础

人体解剖学规定,人体有 3 个互相垂直的基本轴和基本面[11],其在描述身体和关节运动时非常重要。图 2 中 3 个解剖平面分别称为矢状面、冠状面和横断面。

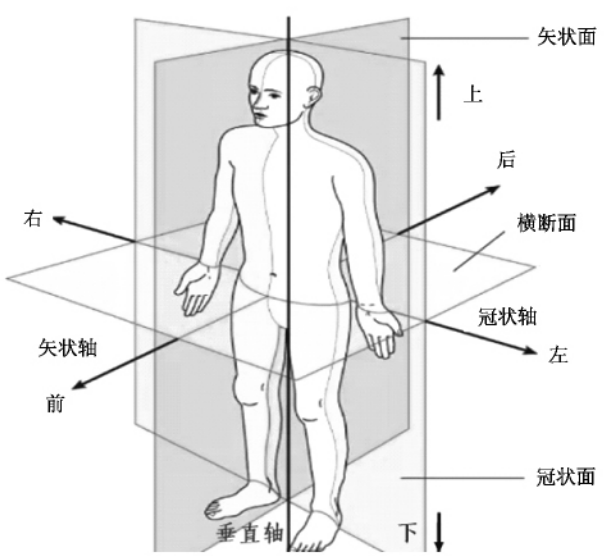

图 2 人体解剖学平面

当描述一个人体的动作时,可以根据占主导作用的平面将动作进行分解。正是基于这种生物学解剖平面的理论基础,将 Kinect V2 捕捉到的骨架数据投影到 3 个解剖平面上,通过计算不同平面上的骨骼轨迹来使得特征表示方法具有时间和空间的特性,并且能够代表所有的人体运动姿态。

通常生物学中冠状面是由人体躯干表示[12],可以采取类似的方式,使用人体躯干关节点来估计冠状面。假设已经获得了 M 个躯干节点的坐标,记为 $P = \{(x_i, y_i, z_i)\}_{i=1}^{M}$,要找到一个最合适躯干节点的平面 $z = Ax + By + C$,就需要估计出 A、B、C 的值,采用最小二乘法可以很好地解决这个问题,当 A、B、C 的误差平方和最小时得到的就是最适合的平面。如公式误差平方和的定义,$err(A,B,C) = \sum_{i=1}^{M} \| (Ax_i + By_i + C) - z_i \|^2$。

2.2 特征向量

2.2.1 构建 3 个投影平面

假设此时已经得到了图 2 中的 3 个解剖学平面,并且以这 3 个解剖平面建立一个坐标系,坐标轴分别为 x_a, y_a, z_a,有某一骨骼点的坐标 $p = (x, y, z)$,那么可以将 t 时刻的这个点分别投影到 3 个平面上,得到 $p_t^{(x_a, y_a)}$、$p_t^{(y_a, z_a)}$、$p_t^{(z_a, x_a)}$,其中 $p_t^{(x_a, y_a)}$ 代表了冠状面坐标,$p_t^{(y_a, z_a)}$ 代表了矢状面坐标,$p_t^{(z_a, x_a)}$ 代表了横平面坐标。采用生物解剖学的投影平面,使得特征表示方法具有旋转不变性和平移不变性[13]。

2.2.2 骨骼空间位置模型

在所有的 3 维骨骼坐标都被投影到 2 维平面上之后,可以选择相邻时间上的运动矢量之间的夹角直方图来表示每个骨骼节点的运动轨迹。先构建骨骼静态位置模型,它表示在静止状态下,人体某一骨骼节点相对于另外节点的位置。如挥手时,人的下肢是相对静止的,而在踢腿时上肢则是相对静止的,可知挥手和踢腿时的手臂相对于躯干中心的位置是不一样的,所以可以明显地区别 2 个动作,这样以躯干中心为原点建立骨骼静态位置模型 $p = \{x_i - x_0 \mid i = 1,2,\cdots,N\}$,其中 x_0 代表躯干中心,N 表示共有多少个节点,实验中 $N = 14$。

2.2.3 计算骨骼角度

骨骼的角度信息不仅代表了节点之间的位置,还表示了人体运动幅度的大小变化。假设有一组骨骼节点数据在任意 2 维平面的坐标 $P = \{p_t\}_1^T$,那么计算相邻时刻 2 坐标之间的夹角的公式为

$$\theta_t = \arccos \frac{\overrightarrow{p_{t-1}p_t} \cdot \overrightarrow{p_tp_{t+1}}}{\| \overrightarrow{p_{t-1}p_t} \| \| \overrightarrow{p_tp_{t+1}} \|}, t = 2, \cdots, T-1 \tag{1}$$

其中 θ 的取值范围是(-180°,180°)。在进行直方图编码时,需要对角度范围进行划分,将(-180°,180°)分为 12 份,即每 30°为一份。计算不同时刻相邻节点的角度,然后统计出直方图对时间特征进行编码。

图 3 中横坐标依次代表 xy,zx,yz 平面上左臂、右臂、左腿、右腿和躯干分为 0~30°、30~60°,…,330~360°的角度值,纵坐标为视频中角度值变化的次数,角度值的变化描述了人体不同关节点运动幅度的大小,代替了以往用骨骼坐标来描述人体动作的方法。而角度的计算方式不受人体身材尺寸大小的影响,即某一骨骼节点向量与另一骨骼节点向量间的夹角,与这两个向量的长短无关,即骨骼节点在同一方向移动的远近不影响角度变化,所以基于方向变化的特征表示方式适用于不同身材尺寸的人,也就是具有不变性。

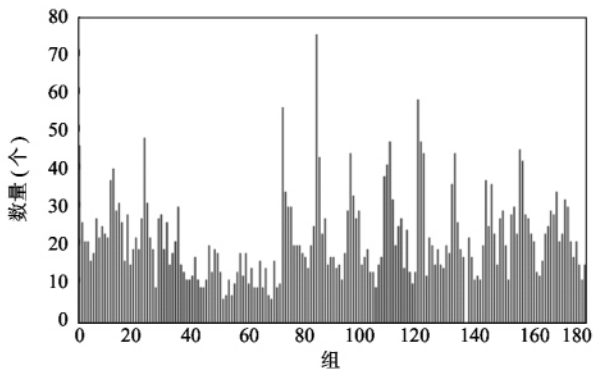

图 3 直方图

2.2.4 基于时间金字塔的特征提取

为了对长时间间隔的时间信息进行编码,使用基于图像金字塔的分维融合算法会是更好的选择。将整段时间内骨骼轨迹分解成 3 个不同等级的图像序列。第 1 等级,将整段时间内的骨骼节点角度变

化全部记录下来，按式(1)选取相邻时间间隔的骨骼节点计算角度变化。第 2 等级，只选取一半时间的骨骼节点坐标，即 $t = 1,3,5,\cdots,n$。第 3 等级，选择更大时间间隔的骨骼节点坐标，即 $t = 1,5,9,\cdots,n$。图 4 中给出了一个直观的时间金字塔表示的实例[14]。

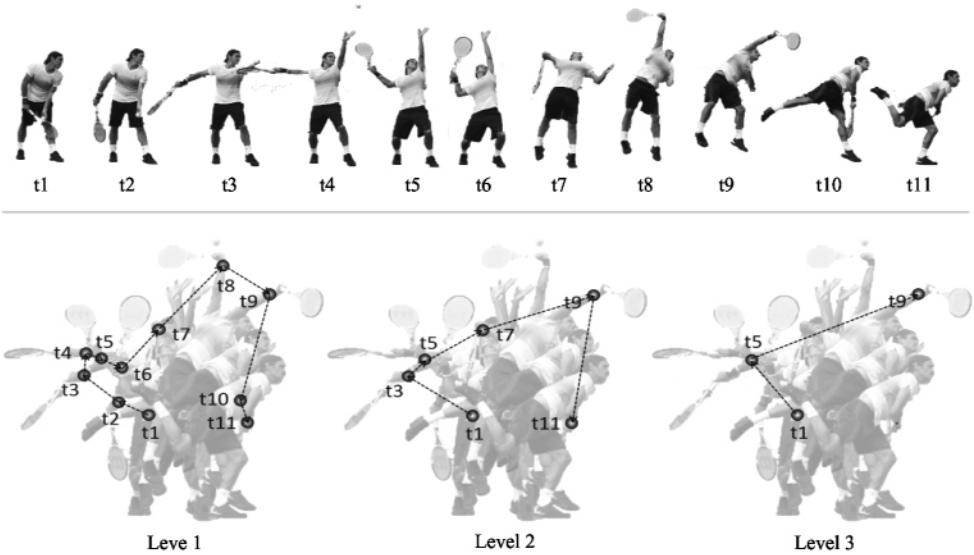

图 4　时间金字塔分解图

在用直方图表示角度变化时，采取将 3 个等级计算出的角度变化累加的方式，即投影到 3 个解剖平面坐标，角度从 0 °到 360 °按每 30 °划分，分为 12 份，还要分别计算 15 个骨骼节点。这种直方图统计角度的特征表示方法会带来很大的计算量，有可能会影响到系统的实时性。当人做某些动作时身体的某些关节并不会产生巨大变化，或者是产生的变化基本一致。举例来说，当人举起双手时，下半身并不会运动，并且手掌和手腕的运动轨迹非常相似。可以通过去除这些冗余的骨骼节点信息，达到减少特征向量维数的目的。根据 Jiang 等人[15]的分析结果，将人体划分为 5 大部分即可以清楚地表示出人体动作。第 1 部分是左臂，包括了左肘和左手掌 2 个骨骼节点。第 2 部分是右臂，包括了右肘和右手掌。第 3 部分是左腿，包括了左膝和左脚。第 4 部分是右腿，包括了右膝和右脚。第 5 部分是躯干，包括了头、脖子、两肩和两胯。经过筛选具有有用信息的骨骼节点，我们最终得到了一个 180 维的特征表示向量，如图 5 所示。

```
In [10]: final_feature_vectors
Out[10]:
array([280,  84,  49,  48,  34,  21,  27,  22,  24,  50,  65,  94, 269,
        91,  68,  40,  17,  29,  31,  23,  31,  41,  67,  91, 177,  90,
        56,  56,  49,  48,  37,  33,  32,  36,  26,  20, 232,  91,  58,
        42,  29,  28,  39,  42,  37,  50,  59,  91, 221,  90,  48,  39,
        53,  42,  40,  30,  40,  54,  71,  70, 190,  96,  40,  43,  38,
        55,  46,  29,  30,  28,  31,  54, 297,  79,  41,  24,  30,  28,
        16,  28,  31,  42,  61, 121, 285,  86,  45,  33,  34,  17,  23,
        23,  45,  44,  60, 103, 199,  84,  53,  45,  40,  53,  35,  26,
        25,  34,  43,  46, 269, 100,  46,  28,  22,  29,  23,  23,  32,
        52,  78,  96, 260,  96,  52,  33,  32,  21,  24,  30,  33,  49,
        79,  89, 196,  78,  52,  51,  39,  39,  37,  33,  32,  23,  37,
        50, 864, 227, 132,  82,  77,  59,  53,  75,  88, 120, 251, 366,
       853, 255, 127,  92,  61,  60,  62,  56, 102, 124, 224, 378, 537,
       240, 161, 132, 111, 134, 114,  97, 105,  97, 119, 129])
```

图 5　180 维特征向量

2.2.5　特征向量改进

在前面的实验方法中，由于获取视频数据时，采取等时间连续存储 5 个子动作的方式，因此每个时间段内的子动作信息量不一致。可能会出现第 1 个时间段内做了 2 次动作，但第 2 个时间段内没有做动作，因此会造成视频数据的不准确。虽然这种方式获取简单，但实际效果不尽如人意。这里借鉴 Yang[16]的自适应能量方法。

首先，利用下列公式计算整个视频数据的总能量，$D_v^n = \sum_{i=1}^{n} | I_v^i - I_v^{i-1} |$，再利用公式 $D_v^i / D_v^n = k/N$。其中，n 表示整个视频序列的帧数，N 表示子动作个数，本文中取 $N = 5$。当 $k = 1$ 时，求得的 i 值即是第 1 个子动作所取的索引帧值，同理 $k = 2$ 时，求得的 i 值即是第 2 个子动作的索引帧值，按照此方法可以根据等能量的方法划分子动作，这样保证了每个部分的动作能量都是相等的。

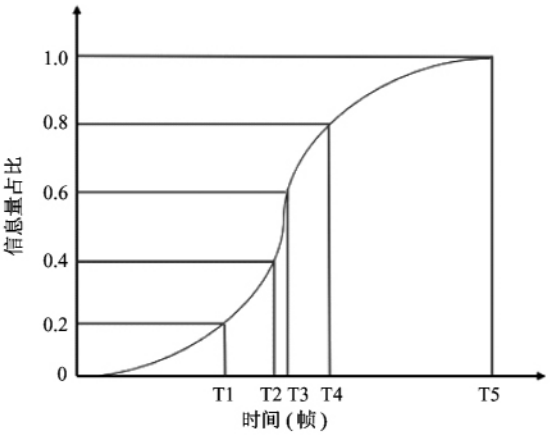

图 6　自适应能量图

2.3　姿势识别

支持向量机(support vector machine,SVM)是目前最常用且效果最好的分类器之一,其具有优秀的泛化能力,对数据规模和分布的要求较低。而且能够解决小样本、非线性和高纬度的问题。并且 SVM 中使用核函数解决特征向量映射到高维后要进行内积的计算,降低了计算的复杂度。传统的 SVM 只针对 2 分类问题,而本文使用的 libsvm 提供的 SVM 库具有多分类的能力[17]。

2.4　骨骼节点预测

因为从 Kinect 中获取的骨骼节点坐标可能出现误差,如图 7 中,当人体出现遮挡情况时,Kinect 不能正确判断骨骼节点的位置,会使得骨骼节点出现漂移,如图中方框圈出的骨骼节点。

图 7　骨骼节点预测

要从已经获得的数据中估计出出错的骨骼节点的坐标,使用扩展卡尔曼滤波(EKF)来估计骨骼坐标。EKF 有 2 个主要优点。一是它符合具有预测能力的要求,EKF 通过迭代前一时刻的状态值和当前时刻的状态值来预测未来时刻的状态值。第二点是由于其迭代的特性,仅需要前一时刻和当前时刻的数据,保证了在获取骨骼数据的帧之间的时间间隔时期也可以正常运行,即用于选取第 2、第 3 等级骨骼节点坐标。这意味着 EKF 可以应用于实际的机器人系统,对于骨骼节点出现漂移的情况,将此时的骨骼节点坐标视为缺失值,通过 EKF 估计出的值来填补该时间点的值,再计算修正后的特征向量,提升了特征向量描述动作的准确性。

3　测试结果

骨骼特征的提取方法使用 Python 实现,而获取彩色视频和骨骼图像使用 C + + 编写,在 ROS 中可以将这 2 部分作为 2 个节点,建立一个话题进行连接。

为了测试特征向量在人体姿态识别中的效果,实验室用多名实验者录制了多组实验视频。录制视频要求每段视频的时间长度为 3 s,共包括 4 种动作:站立、下蹲、挥动双臂和跳。现有的公开人体运动数据集内的动作通常只做一次,例如 MSR Daily Activity 3D dataset 等。而本实验的实验者在 3 s 内做动作的次数可以是一次或多次。这就意味着整段视频内可能包含着不完整的动作,因此带来了动作识别的困难,以往的模板匹配法用在此视频中就达不到很好的效果。图 8 中是累计柱状图,可以看出相同动作间的整体趋势是呈类似的,而不同动作间如挥动双臂和跳的趋势就可以明显看出不同,这证明了次方法提取的特征向量是有效的。根据累计柱状图中不同身体部分的占比,可以确定出哪些部分对动作识别的效果影响更大,从而可以通过设置不同权值的方式,减小带来误判断的身体动作的权值。例如“挥动双臂”动作中身体躯干部份的占比明显小于其他部分,但实际动作中“站”和“挥动双臂”的躯干部分变化应该基本相同,因此减小这部分的权

值，不仅不会影响动作识别的准确性，相反还可以避免类似动作间的误判段。

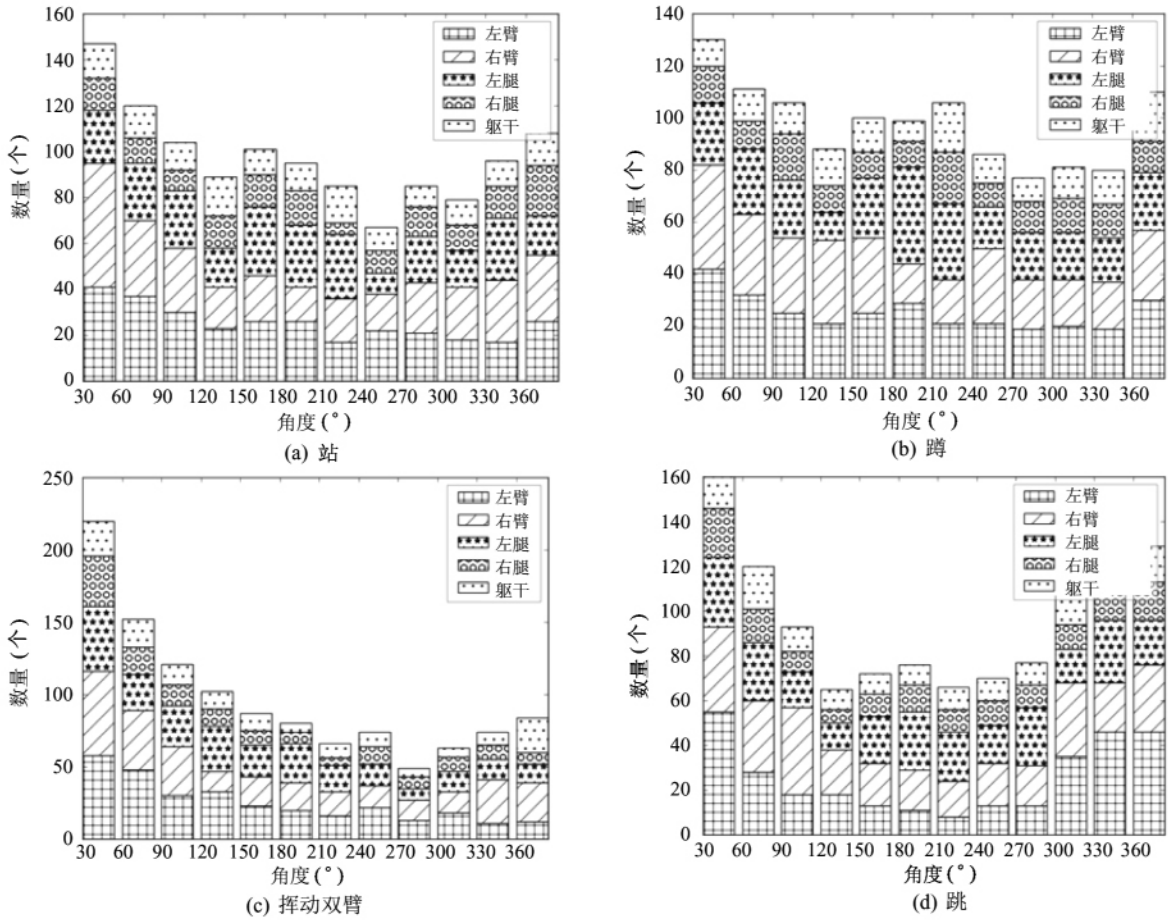

图 8　累计柱状图

最终在调试好 SVM 的参数后，使用了一共 500 组的数据作为训练数据，然后用 10 组数据做验证，得到的交叉验证正确率是 86.25%。通过交叉验证，得到了使得 SVM 分类器效果最好的惩罚系数 $\log_2(C) = -5$ 和 $\log_2(\text{gamma}) = -7$，保证了良好的拟合和泛化能力，以及较快的训练和预测速度。在后面的训练和测试中可以指定惩罚系数和 gamma 值以达到更好的预测效果。

在验证集实验后，得到了如图 9 的混淆矩阵，相比交叉验证的正确率下降了少许，正确率为85.25%。从图中可以看出出现误判率较高的动作为跳，尤其是将跳误判为站。

在采用了自适应能量的方法后再次进行验证，通过等能量划分子动作，可以获得更好的分类效果，图10中的正确率提升到了92.25%。4类动作相比之前的识别率都有提高，其中站、蹲、挥动双臂的正确率均达到了 90% 以上。

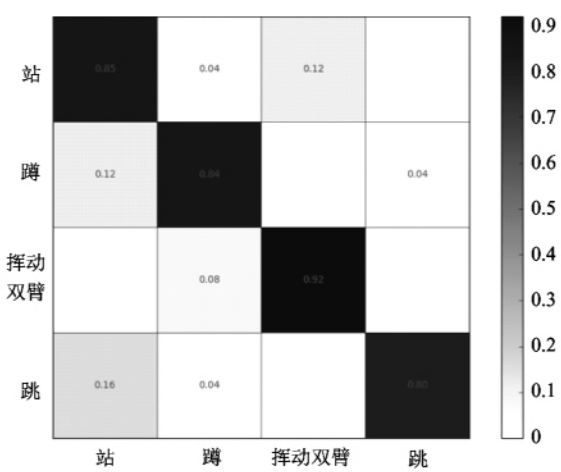

图 9　混淆矩阵

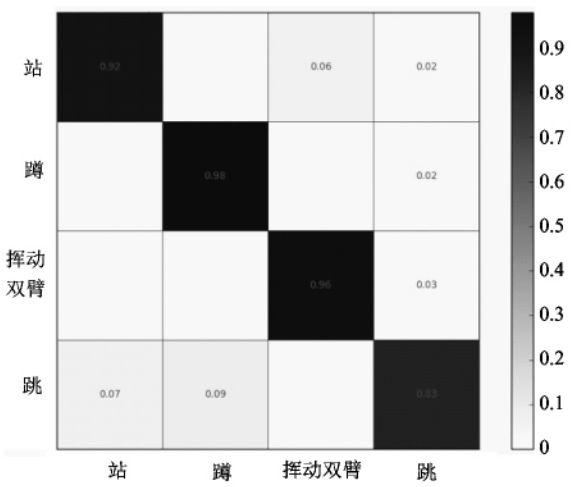

图 10　修改后的混淆矩阵

4　结　论

根据人体解剖学建立的 3 个坐标平面,利用 OpenNI2 提供的骨骼节点坐标建立空间位置模型,计算得到骨骼向量变化的角度。为了对不同时间间隔的骨骼数据进行编码,采取时间金字塔的多级融合算法,将骨骼向量角度分级计算并累加,得到 180 维的具有旋转、平移不变性的特征向量。在此基础上又改进了特征向量,以自适应能量的方法,将一段视频内骨骼角度变化的多少作为衡量能量的标准,按照能量相等划分视频中的人体动作。为了解决人体被遮挡时出现骨骼节点漂移的现象,利用 EKF 预测真实骨骼节点坐标,有效提高了特征向量描述人体动作的准确性。针对身体不同部分骨骼角度变化大小在动作识别中占比不同的情况,设置了不同的权值,减少了冗余骨骼节点带来的误判,同时降低了计算的复杂度。经过实验仿真验证了本文提出方法的有效性。

参考文献

[1] 郑莉莉,黄鲜萍,梁荣华. 基于支持向量机的人体姿态识别[J]. 浙江工业大学学报,2012,40(6):670-675

[2] Silva G, Mello M, Shimabukuro Y, et al. Multitemporal classification of natural vegetation cover in Brazilian Cerrado[C]//2011 6th International Workshop on the Analysis of Multi-Temporal Remote Sensing Images, Trento, Italy,2011: 117-120

[3] Shotton J, Ftzgibbon A, Cok M. Real-time human pose recognition in parts from single depth images[C]//IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011: 1297-1304

[4] Cao L, Ou Y, Yu P S. Coupled behavior analysis with applications[J]. *IEEE Transactions on Knowledge & Data Engineering*, 2012,24(8):1378-1392

[5] Aggarwal J K, Xia L. Human activity recognition from 3D data: A review[J]. *Pattern Recognition Letters*, 2014, 48(1):70-80

[6] Chen G, Giuliani M, Clarke D, et al. Action recognition using ensemble weighted multi-instance learning[C]// IEEE International Conference on Robotics and Automation, Hong Kong, China: 2014: 4520-4525

[7] Pieropan A, Salvi G, Pauwels K, et al. Audio-visual classification and detection of human manipulation actions [C]//Internation Conference on Intelligent Robots and System, Chicago, USA, 2014: 3045-3052

[8] Zhang H, Parker L E. 4-dimensional local spatio-temporal features for human activity recognition[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, USA, 2011: 2044-2049

[9] Yuan J, Wu Y, Liu Z, et al. Mining actionlet ensemble for action recognition with depth cameras[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 1290-1297

[10] Liu Z. Predicting human activities using spatio-temporal structure of interest points[J]. *Multimedia (ACMMM)*, 2012, DOI: 10.1145/239334/2396380

[11] Kersey R D. Color atlas of anatomy: a photographic study of the human body[J]. *Springhouse Pub Co*, 1987, 89(4):2740

[12] McGinnis M. Bioregionalism: The Tug and Pull of Place [M]. New York: Routledge, 1999: 183-186

[13] 项海兵,刘劲松,吴涛,等. 机载 SAR 图像的动态金字塔实时显示技术[J]. 中国图像图形学报, 2018,23(12):1938-1946

[14] Zhang H, Parker L E. Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction[C]//2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, USA, 2015: 3053-3060

[15] Jiang M, Kong J, Bebis G, et al. Informative joints

based human action recognition using skeleton contexts [J]. *Signal Processing Image Communication*, 2015, 33 (C):29-40

[16] Yang X, Tian Y L. Super normal vector for activity recognition using depth sequences [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, 2014: 804-811

[17] Chang C C, Lin C J. LIBSVM: A library for support vector machines [J]. *ACM Transactions on Intelligent System and Technology*, 2011, 2(3):1-27

Research on human body attitude recognition based on Kinect and ROS

Hu Dunli, Ke Haoran, Zhang Wei

(Beijing Key Laboratory of Fieldbus and Automation, North China University of Technology, Beijing 100049)

Abstract

In order to solve the problem of incomplete human motion recognition, the angular cumulative change of the skeleton node vector is proposed as the feature vector under the robot operating system (ROS), and the adaptive human energy method is used to divide the video of human body motion. The projection coordinate plane is established on the basis of human anatomy, and the calculation of the spatial position and the bone angle is performed. The time pyramid method is used to encode the bone angle data of different time intervals, and the multi-level feature vector is formed to better represent the human body motion. In the case of human occlusion, the extended Kalman filter is used to predict the coordinates of the bone nodes and improve the accuracy of the bone coordinates. The method has rotation and translation invariance, and the correct rate of identifying 4 incomplete human movements reaches 92.25%.

Key words: adaptive energy, time pyramid, extended Kalman filter, robot operating system (ROS), prediction