doi:10.3772/j.issn.1002-0470.2019.09.005

基于规则的连续动作识别①

丁伟利② 胡 博 张焱鑫

(燕山大学电气工程学院 秦皇岛 066004)

摘 要 动作识别技术成为近些年计算机视觉领域的研究热点。本文针对深度传感器获得的骨骼信息,提出一种基于规则的动作识别算法。首先设定规则定位动作,将连续动作分段划分,使整段动作划分成短的多段动作数据。将视频中所获得的不同动作的深度骨骼数据用不同方法进行归一化。通过动态时间规整(DTW)算法对分段动作数据进行分析,得到最匹配的动作标签,实现动作识别。从实验结果可知,该方法在动作识别中具有较高的准确性以及较好的实用性。

关键词 动作识别,规则,分段,动态时间规整(DTW)算法

0 引 言

近些年,动作识别技术在人机交互、康复医疗[1]、事件监测、虚拟现实等领域变得日益重要,成为计算机视觉领域的研究热点。动作识别技术主要是指计算机对于视频、图像或者 RGD-B 图像序列中的动作行为进行识别和对比。通过识别算法使机器设备更好地理解人类的动作语言,通过肢体动作引导机器设备进行生产工作。尤其是在嘈杂的生产车间及危险的工作环境中,动作识别这种人机交互方式相比按键控制、语音交互等方式有着更为准确、高效和自然的优势。

早期动作识别的研究方法多通过单摄像头 2D 图像提取运动特征。James 等人[2]采用运动能量模型(motion-energy image,MEI)和运动历史模型(motion-history image,MHI)对人体动作特征进行表征。但是由于视频拍摄的图像,受衣着、光照、遮挡以及拍摄角度不同的影响较大,同时很难排除复杂背景的干扰,因此基于 RGB 视频获得特征方法的局限性也渐渐显露出来。近几年,随着深度图像传感器成本越来越低,基于 RGD-B 图像的动作识别技术逐渐成为动作识别研究领域的一个重要方向,研究者利用 Kinect 获取的骨骼数据提出了众多动作识别算法,且准确性得到了大幅提升。

目前,基于骨骼数据的连续人体动作识别方法主要有以下几种方式。第 1 种是基于概率统计的方法,其中最常用到的模型有动态贝叶斯网络(dynamic Bayesian network, DBN)、隐马尔科夫模型(hidden Markov models, HMM)以及支持向量机(SVM)3 种方法。Yang 等人[3]基于朴素贝叶斯和最邻近两种方法实现动作识别。Yamato[4]和 Hongeng[5]基于隐马尔科夫模型的人体动作识别方法进行研究。支持向量机(SVM)[6,7]方法是使用该向量机学习训练样本,将测试样本分类,最终获得动作识别结果。第 2 种是基于模板的方法,可分为两种,一种是帧对帧的匹配方法,较为常用的是动态时间规整(dynamic time warping, DTW)算法[8],将动态时间规整算法应用到人体动作识别中,能够将待识别的动作特征与模板中每个时刻特征进行匹配,计算出累计距离,获得两条动作曲线的最小的匹配路径。文献[9]提出了模板匹配方法,先建立标准动作模

① 国家自然科学基金(61573356)和河北省自然科学基金(F2016203211)资助项目。

② 女,1979 年生,博士,教授;研究方向:模式识别,虚拟现实,计算机视觉;联系人,E-mail: weiye51@ysu.edu.cn

(收稿日期:2018-11-03)

板库,然后将采集的动作数据与模板数据对比匹配,进行动作识别。动态规划算法[10]是将待识别的样本模板中每个时刻的特征与特征模板中的任意时刻进行特征匹配。近些年,被广泛关注的深度学习方法[11,12]也渐渐推广到动作识别中来,虽然有较好的识别结果,但深度学习生成的模型较难解释,在实验过程中通常需要较大的数据集,同时在参数调节方面也需要较多时间。

虽然近些年人体动作识别取得了很好的成果,但同时还有很多不足和挑战。第 1 方面是人体的结构与动作,不同的人身高不同、体态不同、动作幅度也不相同,因此给动作识别的准确性增加了难度。第 2 方面是人体动作的切割,由于动作是连续的,无法确定何时开始及何时结束,使得识别有很大的难度。第 3 方面是人体动作数据的高维表示,以 Kinect 获得的动作数据为例,每组连续数据至少是 20 ×3 =60 维的人体骨骼数据,因此,动作识别算法的实时高效性也变得尤为重要。

针对现有动作识别的挑战性问题,本文主要基于 Kinect 获得运动骨骼数据,提出了基于规则动作识别方法。首先以 Kinect 获得的连续动作数据作为输入,然后定义多种规则定位开始动作帧与结束动作帧,使用开始及结束帧将连续动作分段。因为动作数据在未分段前数据量过大,且每个动作反复执行次数不确定,开始结束时间点也不确定,动作分段能将整段数据划分成短的多段动作进行识别,在提高准确率的同时也获得了更丰富的反馈信息。最后,分段后得到的动作数据经归一化后结合 DTW 算法与模板库数据对比识别,获得具体动作标签。

1　动作定位和自动分段

动作识别与语音识别相似,每个动作都是由多个连续帧组成的,动作与动作间没有明确的界限,因此,若能很好地将连续动作分段,将会大大降低动作识别的难度。本文提出了基于规则的动作定位识别算法。其中,定义两个方面规则,一方面是骨骼段夹角,另一方面是运动姿态判断。在动作曲线分段前,首先判断是否伴随下肢动作(如弯腰、下蹲等),如有伴随下肢动作,经过分段后的动作曲线仅需与有伴随下肢动作的模板匹配,无需与全部模板一一匹配,可减少识别过程的计算量。骨骼夹角特征的定义可以从人体动作相关结构上很好地描述动作,同时针对不同身高、不同身材的人,所定义的三维空间的角度特征具有旋转及尺度不变性。

由于 Kinect 传感器可以实时获得人体 20 个关节点的 3 维位置信息,被主动跟踪的用户关节点信息由 x、y、z 坐标表示,单位为 m。关节空间坐标信息由 Kinect 位置决定,如果 Kinect 放置位置不是水平面,或者 Kinect 摄像头与传动马达之间有夹角,那么被采集的用户就算是笔直站立的,采集的数据也是倾斜的,Kinect 关节空间原点坐标为 Kinect 位置,空间坐标如图 1 所示。

Kinect 获得的关节点坐标 J 数量一共有 20 个,每个关节点在 t 帧时刻的坐标为 $p_i(t)$,其中:

$$p_i(t) = (x_i(t), y_i(t), z_i(t)); (i \in [1, 20]) \tag{1}$$

任意两个关节点 i, j 组成的骨骼段直线方程的方向向量 $\boldsymbol{s}_m(\boldsymbol{a}_m, \boldsymbol{b}_m, \boldsymbol{c}_m)$ 为

$$\boldsymbol{s}_m(\boldsymbol{a}_m, \boldsymbol{b}_m, \boldsymbol{c}_m) = (x_j(t) - x_i(t), y_j(t) - y_i(t), z_j(t) - z_i(t)) \tag{2}$$

两骨骼段夹角 $Angle$ 为

$$Angle = \arccos((\boldsymbol{a}_1 \cdot \boldsymbol{a}_2 + \boldsymbol{b}_1 \cdot \boldsymbol{b}_2 + \boldsymbol{c}_1 \cdot \boldsymbol{c}_2) / (\sqrt{\boldsymbol{a}_1^2 + \boldsymbol{b}_1^2 + \boldsymbol{c}_1^2} \cdot \sqrt{\boldsymbol{a}_2^2 + \boldsymbol{b}_2^2 + \boldsymbol{c}_2^2})) \tag{3}$$

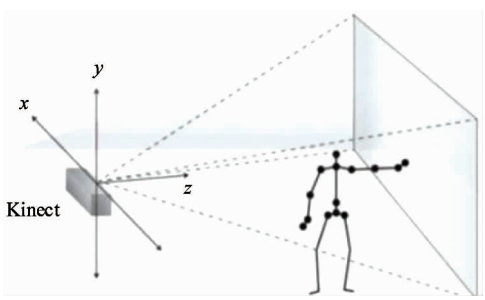

图 1　Kinect 采集空间坐标系

为了确保特征的完备性,本文一共定义 23 个骨骼段,两两骨骼段取夹角后,共得到 12 个波动较为明显的夹角特征。在参与者执行动作时,左臂运动时波动较为明显的共有 3 个夹角;右臂运动时波动较为明显的共有 3 个夹角;左腿运动时波动较为明

显的共有 3 个夹角;右腿运动时波动较为明显的共有 3 个夹角。由于这 12 个动作波动较为明显,可以较为准确地区分不同动作,因此选取这 12 个特征角度(见表 1)来定义不同动作的开始帧和结束帧,用开始帧和结束帧来划分连续动作,使用划分好的动作数据进行动作识别时结果会更为准确。

表 1　动作分段特征夹角

动作类型	波动明显的相关动作夹角	夹角标号
	右肩到左肩骨骼段与左肩到左肘骨骼段夹角	1
左臂动作	左肩到左肘骨骼段与左肘到左手腕骨骼段夹角	2
	左肩到左肘骨骼段与肩部中心到髋部中心骨骼段的夹角	3
	右肩到左肩骨骼段与右肩到右肘骨骼段夹角	4
右臂动作	右肩到右肘骨骼段与右肘到右手腕骨骼段夹角	5
	右肩到右肘骨骼段与肩部中心到髋部中心骨骼段的夹角	6
	左髋到左膝骨骼段与肩部中心到髋部中心的夹角	7
左腿动作	左髋到左膝骨骼段与左膝到左脚踝骨骼段夹角	8
	左髋到右髋骨骼段与左髋到左膝骨骼段夹角	9
	右髋到右膝骨骼段与肩部中心到髋部中心骨骼段的夹角	10
右腿动作	右髋到右膝骨骼段与右膝到右脚踝骨骼段夹角	11
	左髋到右髋骨骼段与右髋到右膝骨骼段夹角	12

首先对动作数据进行“姿势规划”,目的是判断动作过程中是否伴随下肢动作。如果该动作未伴随有下肢动作,则本文所定义的 6 个左右腿夹角在整段动作夹角曲线中无明显变化,当有明显变化时可判断该动作有伴随下肢动作。图 2 为左腿动作及右腿动作定义的共 6 个夹角在无伴随下肢动作时的夹角变化曲线,图3则为有伴随下肢动作的6个夹角变化曲线。由图中可见,当该动作无伴随下肢运动时,6 个夹角几乎没有变化,当动作有伴随下肢运动时 ,6个夹角均有明显波动。在图2与图3中,横坐标轴为该动作序列的帧数 F,纵轴为夹角度数 $Angle$。本文设定了针对无伴随下肢动作时 6 个腿部动作角度的变化范围,当全部动作序列帧数中 2% 的帧数夹角角度超过规定的范围,则认为该动作序列有伴随下肢动作。

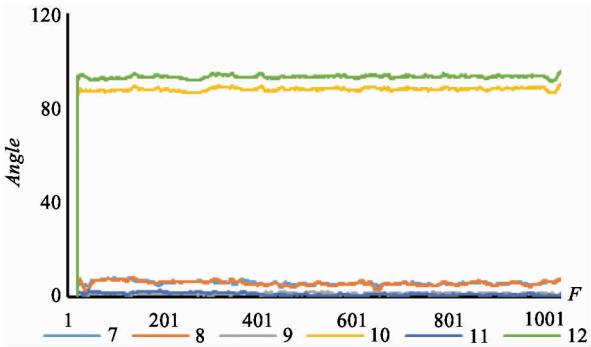

图 2　无伴随下肢动作时的夹角变化曲线

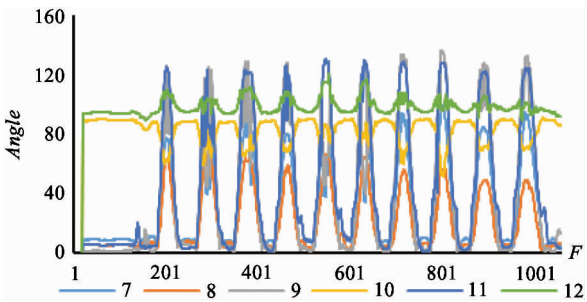

图 3　有伴随下肢动作时的夹角变化曲线

姿态划分后,定位开始帧和结束帧动作将连续动作中有效动作分割出来。如一组连续序列帧集合 $F = \{f_1, f_2, \cdots, f_n\}$;当前帧动作特征满足开始帧动作规则时,标记该帧为开始帧 f_{si},开始帧集合 $F_{start} = \{f_{s1}, f_{s2}, \cdots, f_{si}, \cdots, f_{sk}\}$,($F_{start} \in F$, $i \in [1, k]$);当前帧动作特征满足结束帧动作规则时,标记

该帧为结束帧 f_{ei}，结束帧动作集合 $F_{end} = \{f_{e1}, f_{e2}, \cdots, f_{ei}, \cdots, f_{ek}\}$，($F_{end} \in F, f_{si} < f_{ei}, i \in [1, k]$)，得知开始帧及结束帧位置后，组成连续动作曲线中的有效动作帧集合 $F_{Ef} = \{f_{ef1}, f_{ef2}, \cdots, f_{efi}, \cdots, f_{efk}\}$，($f_{efi} \in [f_{si}, f_{ei}], i \in [1, k], E_{Ef} \in F$) 将有效动作分割出来进行动作识别。因此采用本文提出的动作定位与自动分段识别算法，无需识别全部动作曲线，仅自动提取出有效动作曲线完成分类识别。同时该方法的动作定位及自动分段功能可以获得连续动作的开始帧和结束帧的帧数位置以及动作重复次数等更为丰富的信息。

2 动作识别

DTW 算法最初由日本学者 Itkura[13] 提出，主要用于比较两个时间序列的相似性，早期应用于语音识别[14]领域。在语音识别中，读同一个单词的语速快慢不同，但音调是相似的，动作识别与之类似，虽然参与者所做动作快慢不同，但是运动轨迹是相似的。如何将两段长度不同的时间序列曲线进行拉伸或压缩，使两段时间序列曲线在具有相同长度的同时也具有较好的匹配度，是实现动作识别的关键。如图 4 所示，两个相似时间序列曲线进行匹配时，按照传统的欧几里得定理，a 点对应的是 c 点，但实际 a 点对应的是 b 点，运用 DTW 算法就能很好地解决这一问题。

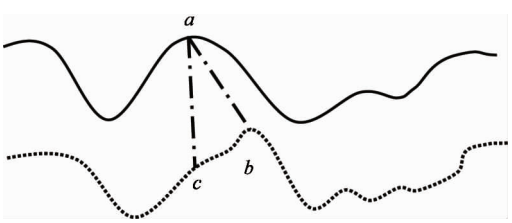

图 4 DTW 算法与传统匹配方法的区别

然而在实际中采集动作时，参与者所站立的位置及所面对的方向不同，并且在动作过程中位置也是在实时变化的，因此采集到的动作数据差异较大。在动作识别时，计算两条运动轨迹相似性会产生很大偏差，为了保证采集的动作数据的统一性，需要将数据进行归一化。本文主要针对两种动作姿态采用两种不同的归一化方法，除下蹲姿势外我们将所有动作的每一帧除髋部中心的其他关节点归一化到当前帧髋部中心位置，将以 Kinect 所在的空间位置为坐标原点的原坐标系转化为以髋部中心空间位置为原点的新坐标系。假设当前帧髋部中心原始坐标为 (x_0, y_0, z_0)，以头部坐标为例，头部的初始坐标为 (x_1, y_1, z_1)，那么归一化后，髋部中心坐标为 $(x_0', y_0', z_0') = (0, 0, 0)$，头部坐标 $(x_1', y_1', z_1') = (x_1 - x_0, y_1 - y_0, z_1 - z_0)$，其他关节点的归一化坐标同样根据该方法计算可得。针对下蹲动作进行模板匹配时，由于下蹲动作的髋部中心坐标实时变化波动较大，所以在动作识别中，髋部中心运动轨迹曲线在相似性度量时占据重要的地位，不能按上述归一化方法将髋部中心归一化到原点。因此本文采用的方法是将每个下蹲动作刚开始位置定义为原点，然后将每个下蹲动作分别归一化到每个下蹲开始动作的原点位置。

将骨骼数据归一化后，本文将分割好的动作曲线结合 DTW 算法进行动作曲线相似性判断。具体方法如下：

假设模板的第 i 个关节点坐标为 (x_i, y_i, z_i)，待识别的动作数据第 i 个关节点坐标为 (x_i', y_i', z_i')，则相似性距离 d 的计算公式为

$$d = (x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2 \quad i \in [1, 20] \quad (4)$$

根据式(4)可知，归一化可以减小原始坐标空间的相似距离误差大的情况。本文选择了动作时波动较为明显的 11 个关节点的运动轨迹计算曲线相似性，分别为左肩、右肩、左肘、右肘、左手腕、右手腕、左膝、右膝、左脚踝、右脚踝和髋部中心。在识别无伴随下肢运动的动作时，将关节点轨迹结合双手间距特征和双肘间距特征以及双手到头距离特征进行识别。采用该方法，在获得动作识别结果的同时也能得到与模板动作的相似性距离 d(式(4))。对相似性距离逐帧累计求和可得到该动作曲线与模板的一一匹配最小累计距离差值，该值越小说明该动作与模板标准动作越相似(与模板完全一致时该值为 0)，通过该值可知该动作与标准模板动作的相似性，因此在动作识别中有更好的实用性。

3 实验结果

本文实验所选择的动作数据集为 MSRC-12(Microsoft Research Cambridge-12)微软剑桥大学计算机实验室采集的人体动作数据集,该数据集由人体 20 个关节点的时间序列数据组成,由 Kohli 等人[15]通过 Kinect 深度传感器训练采集获得。该数据集采集了 30 个参与者,在 5 种方式引导下分别执行 12 种动作获得,这 5 种方式分别为分解动作的描述文本、动作的静态图片、动作视频、文本和视频以及文本和图片。所执行的 12 种动作分为两类,一类是标志性动作,另一类为抽象的隐含类动作。

本算法先规划姿势识别,判断姿态是否包含下肢动作。通过姿势划分以及起始终止动作定位后,对分段好的动作曲线结合 DTW 算法进行识别,动作识别的准确率与其他方法的对比如下:

由图 5 可见,本文算法在识别准确率方面优于传统的 DTW 算法、HMM 算法[16]以及 Veloangles-SVM[17]方法,同时本文算法如表 2 所示,以 MSRC-12 数据集中文件数据作为输入,表中列出每个动作的总样本数以及其中识别正确样本数,由表 2 可见识别正确的样本平均相似性距离值,值越小表明该动作与模板动作越相似。由于 MSRC-12 每个样本动作为重复多次的,本文提出的动作定位及自动分段识别算法在获得该动作分类标签的同时也可获得动作数据开始帧及结束帧的帧数值,表 2 也列出了动作分段的准确度以及相似性距离值。可见在该算法能获得该动作的标签以及与模板动作相比的相似性累加距离的同时也能获得参与者在执行动作时的起始帧、结束帧以及重复该动作次数等更为丰富的信息。由此可见,本文算法相比机器学习方法具有更好的解释性,使用户对判别规则有更直观的了解,因此在如指导肢体动作功能障碍患者进行康复训练等方面具有更好的实用性。

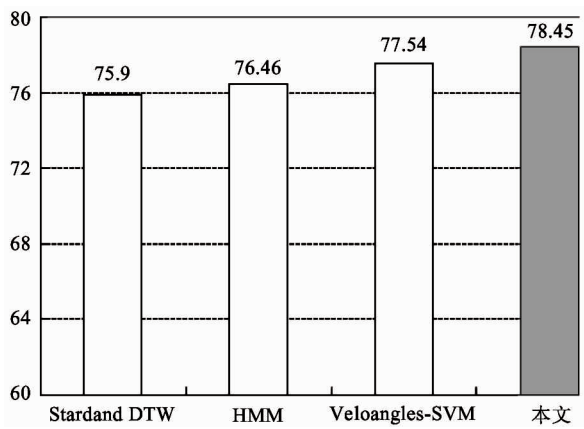

图 5 实验结果对比

表 2 MSRC-12 中每个动作的相似性距离及分段准确度

动作标签	总样本数	识别正确样本数	相似性距离	分段准确度
A1	50	39	0.191263	96.41%
A2	50	44	0.238190	79.77%
A3	50	46	0.276330	58.04%
A4	50	43	0.225362	79.07%
A5	48	22	0.486755	47.27%
A6	49	39	0.220679	64.10%
A7	50	48	0.127975	78.13%
A8	50	37	0.573459	78.11%
A9	50	42	0.230444	80.48%
A10	48	43	0.206715	77.67%
A11	49	18	3.932380	61.11%
A12	50	45	0.477734	56.89%

4 结 论

本文提出了一种基于规则的动作识别算法,首先将连续的动作序列按照定义好的规则进行动作定位与自动分段,其中经过姿势规划规则将动作数据划分为有伴随下肢运动和无伴随下肢运动的两种动

作,骨骼夹角规则的定义保证开始帧和结束帧动作定位的准确性,结合 DTW 算法对分割好的中间动作轨迹进行动作识别。DTW 算法虽然简单,但是针对较长的动作序列计算量非常大,因此运算效率很低,且降低了识别的准确性。此外,本文还针对不同动作姿态采用了不同的归一化方法,更为合理地统一了动作数据。识别中我们选择全部关节点中变化较为明显的 11 个关节点计算关节运动轨迹相似性,在获得每个动作识别结果的同时,能够得到每个关节点运动轨迹与标准模板动作关节点运动轨迹的动作相似度。

因此,本文定义的规则保证了曲线分段的准确性,同时提出的短动作曲线划分策略,相比整段长曲线,在结合 DTW 算法进行识别时的准确度更好,同时该算法获得的信息更为丰富:不但能得到动作分类的标签,也能得到动作的起始与终止帧位置,以及动作的重复次数等信息。识别结束后还可以计算出被试者所做动作与模板动作的相似性,被试者可根据该差值得知自己所做动作与模板动作的差异大小。因此本文提出的算法适合应用于康复和体育动作训练等领域。本研究下一步的工作是进一步提升算法的准确率,并将该算法获得的丰富信息应用在有肢体功能障碍患者的康复动作识别等方面。

参考文献

[1] Ding W L, Zheng Y Z, Su Y P, et al. Kinect-based virtual rehabilitation and evaluation system for upper limb disorders: a case study[J]. *Journal of Back and Musculoskeletal Rehabilitation*, 2018, 31(4): 611-621

[2] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23(3): 257-267

[3] Yang X D, Tian Y L. EigenJoints-based action recognition using Naĭve-Bayes-Nearest-Neighbor[C]. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, USA, 2012. 14-19

[4] Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using hidden Markov model[C]. In: 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Yokosuka, Japan, 1992. 379-385

[5] Hongeng S, Nevatia R, Bremond F. Video-based event recognition: activity representation and probabilistic recognition methods[J]. *Computer Vision and Image Understanding*, 2004, 96(2): 129-162

[6] Qian H, Mao Y, Xiang W, et al. Recognition of human activities using SVM multi-class classifier[J]. *Pattern Recognition Letters*, 2010, 31(2): 100-111

[7] Wu X, Duan L, Duan L, et al. Action recognition using multilevel features and latent structural SVM[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013, 23(8): 1422-1431

[8] Wang J, Zheng H. View-robust action recognition based on temporal self-similarities and dynamic time warping [C]. In: 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE), Zhangjiajie, China, 2012. 498-502

[9] 谢林海, 刘相滨. 基于不变矩特征和神经网络的步态识别[J]. 微计算机信息, 2007, 23(19): 279-281

[10] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. *IEEE Transactions on Acoustics Speech and Signal Processing*, 2003, 26(1): 43-49

[11] Li Y H, Lan C L, Xing J L, et al. Online human action detection using joint classification-regression recurrent neural networks[C]. In: 2016 ECCV Conterence, Amsterdam, Netherlands, 2016. 203-220

[12] Zhu W, Lan C, Xing J, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]. In: 30th AAAI Conference on Artificial Intelligence, Phoenix, USA, 2016. 3697-3703

[13] Itakura F. Minimum prediction residual principle applied to speech recognition[J]. *IEEE Transactions on Acoustics Speech and Signal Processing*, 1975, 23(1): 67-72

[14] Morales-Cordovilla J A, Cabañas-Molero P, Peinado A M, et al. A robust pitch extractor based on DTW lines and CASA with application in noisy speech recognition [C]. In: Communications in Computer and Information Science, Madrid, Spain, 2012. 197-206

[15] Fothergill S, Mentis H, Kohli P, et al. Instructing people for training gestural interactive systems[C]. In: Proceed-

ings of the SIGCHI Conference on Human Foutors in Computing Systems, New York, USA, 2012. 1737-1746

[16] Choi H R, Kim T. Modified dynamic time warping based on direction similarity for fast gesture recognition[J]. *Mathematical Problems in Engineering*, 2018 (2):1-9

[17] Nguyen D D, Le H S. Kinect gesture recognition: SVM vs. RVM[C]. In: International Conference on Knowledge and Systems Engineering, Hochi Minh City, Vietnam, 2016. 395-400

Rule-based continuous action recognition

Ding Weili, Hu Bo, Zhang Yanxin

(School of Electrical Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

Action recognition technology has become a research hotspot in the field of computer vision in recent years. In this paper, a rule-based action recognition algorithm is proposed based on the skeleton information obtained by the depth sensor. Firstly, the regular positioning action is set up, and the continuous action is divided into segments, so that the whole action is divided into short multi-segment action data. Secondly, the depth skeleton information obtained from different actions in the video is normalized by different methods. Finally, the dynamic time warping (DTW) algorithm is applied to analyze the segmented action data. On this basis, the most matching action labels are obtained and action recognition is realized. It can be seen from the experimental results that the method has considerable accuracy and good practicability in action recognition.

Key words: action recognition, rule, segmentation, dynamic time warping (DTW) algorithm