文章编号：1006-4710(2015)03-0253-12

基于深度信息的人体动作识别研究综述

陈万军，张二虎

（西安理工大学 印刷包装与数字媒体学院，陕西 西安 710048）

摘要：随着低成本深度传感器的发明，尤其是微软 Kinect 的出现，高分辨率的深度与视觉(RGB)感知数据被广泛使用，并为解决计算机视觉领域中的基本问题开拓了新的机遇。本文针对基于深度信息的人体动作识别研究，首先提出了一种基于特征和数据类型的分类框架，并对最近几年提出的相关方法进行了全面回顾。随后，对文献中描述的算法进行了性能对比分析，同时对所引用的公共测试数据集进行了总结。最后，笔者对未来的研究方向进行了讨论并给出了相关建议。

关键词：人体动作识别；深度传感器；Kinect；骨架关节点；深度数据

中图分类号：TP391　　**文献标志码**：A

A review for human action recognition based on depth data

CHEN Wanjun, ZHANG Erhu

(Faculty of Printing, Packaging Engineering and Digital Media Technology, Xi'an University of Technology, Xi'an 710048, China)

Abstract: With the invention of the low-cost depth sensors, especially the emergence of Microsoft Kinect, high-resolution depth and visual (RGB) sensing data has become available for widespread use, which opens up new opportunities to solve fundamental problems in computer vision community. This paper presents a comprehensive review of recent depth-based human action recognition algorithms. Firstly, we develop a taxonomic framework according to features and original data type. Following our taxonomy, recent published research on the use of depth data for recognizing human action is reviewed. Then, the publicly available datasets cited in their work are listed. Finally, the authors discuss and suggest future research directions.

Key words: human action recognition; depth sensors; Kinect; skeleton joints; depth data

人类具有从视觉信息感知人体动作的非凡能力，能够定位人、物并跟踪人体运动，通过分析人与物体的交互来理解人们在做什么甚至推断其意图。自动进行人体动作的理解与识别对许多人工智能系统来说是至关重要的，可广泛地应用于智能视频监控、人机交互、视频分析与检索以及机器人等领域中。例如，为了建造一个能智能化地服务于人类的人机交互系统，该系统不仅需要感知人体的运动，而且还要理解人体动作的语义并推断其意图。

基于视觉(RGB)信息的人体动作识别在过去数十年间取得了很大进展[1-2]，许多方法被相继提出。这些方法包括人体关键姿态[3]、运动模板[4]、剪影[5]和时-空形状[6]。最近，时-空兴趣点[7]和密集运动轨迹[8]由于其出色的性能而备受人们的关注。尽管这些方法在相关标准测试数据集上均取得了较好的识别结果，但是由于人体动作具有高度的灵活性，人体的姿态、运动、衣着具有显著的个体差异性，相机视角、相机的运动、光照条件的变化、遮挡、自遮挡并同时包含人-物的交互以及复杂的时-空结构等因素的综合影响，使得人体动作识别仍极具挑战性[1]。

最近，传感器技术的进步为我们带来了低廉的高清晰的深度传感器，例如微软公司的 Kinect 和华硕公司的 Xtion PRO LIVE。深度传感器利用结构光实时地感知世界并形成深度映射(Depth Map)图像。深度映射图像中的每个像素记录着场景的深度值，与普通的 RGB 图像中像素所表示的光强度值截然不同。深度传感器的引入能够极大地拓展计算机系统感知三维世界并提取低层视觉信息的能力，并

收稿日期：2015-03-12

基金项目：国家自然科学基金资助项目(61073092)。

作者简介：陈万军，男，博士生，研究方向为计算机视觉和模式识别。E-mail：wjchen@xaut.edu.cn。

通讯作者：张二虎，男，教授，博导，研究方向为图像处理、模式识别与智能信息处理。E-mail：eh-zhang@xaut.edu.cn。

在计算机视觉领域中受到越来越广泛的应用[9]。2013 年 10 月,IEEE 会刊 Cybernetics 出版了题为“Computer Vision For RGB-D Sensors: Kinect And Its Applications”的专刊,以促进深度传感器在人体动作识别、手势识别及人体姿态估计等热点研究领域的发展。2014 年,国际顶级会议——计算机视觉与模式识别(CVPR 2014)也设立了有关深度相机在解决计算机视觉领域问题的专题研讨会(Towards solving real-world vision problems with RGB-D cameras)。深度传感器较传统 RGB 相机在人体动作识别方面具有无可比拟的优势,即它不受光照条件的影响,具有颜色与纹理不变性,同时深度信息能极大简化目标的检测与分割任务。因此,近来人们提出了各种基于深度传感器的人体动作识别方法。本文依据这些方法所采用的特征和依赖的原始数据类型对其进行分类,具体可分为:基于骨架关节点的方法、基于深度映射图像的方法、基于 3D 点云的方法、RGB 图像与深度信息融合的多模态方法以及基于深度学习的方法。

1 分类方法

1.1 基于骨架关节点的方法

骨架关节点是人体的视觉显著点,其在 4D 空间中的运动能够反映动作的语义信息。基于关节点的动作识别研究可追溯至早期 Johansson 的工作[10]。他们通过实验表明绝大部分动作可仅根据关节点的位置进行识别。这一观点被后续大量的研究者采用并逐渐形成人体动作识别方法的一个重要分支。由于不同时刻的关节点具有自然的对应性,而在传统的 RGB 图像中很难建立起这种对应关系,因此大多数基于关节点的方法利用它们来显示建模人体动作的时间动态特性。目前,人们主要通过 2 种手段来获得关节点的位置信息:基于运动捕捉系统(Motion Capture system,MoCap)[11]和基于深度映射图像的关节点位置估计[10]。需要指出的是,这 2 种方法所获得的关节点位置信息的精度存在一定的差异。MoCap 方法利用多相机并配以关节点标记来进行可靠的关节点位置估计,因而具有更低的嵌入噪声。而后者由于使用单一的深度相机而产生的自遮挡或互遮挡问题,往往使得提取的关节点位置信息具有较低的质量。另外,MoCap 是基于标记点的,只能在特殊的室内环境中使用,且其价格昂贵。早期的基于骨架关节点的动作识别方法多基于此[12-13]。随着 2009 年 Kinect 深度相机的问世,使得关节点位置的估计与跟踪只需一架深度相机,其装置简单,成本低廉,且具有很强的适用性等优点,因而,目前的人体动作识别研究工作更多地集中在基于深度相机的含噪声的关节点数据上。下面,笔者首先介绍基于 MoCap 数据的方法,然后重点讨论基于深度相机的关节点数据方法。

1.1.1 基于 MoCap 的方法

利用运动捕捉系统来采集人体关节点的运动数据需要在特定的场景中进行,并且需要演员穿上特制的带有发光标志的衣服。这些发光标志均位于人体的各个关节点上。多个经标定后的高精度相机阵列环绕排列于场景中用以跟踪并记录这些标志点的 2D 图像,最后经由系统精确地重构出这些标志点(关节点)的 3D 坐标。为了有效刻画 MoCap 数据的本质特征,Müller 等人[12-13]采用关系运动特征(Relational Motion Features)来描述运动姿态的关节点间的几何(语义)关系。关系运动特征具有全局方向、位置、体型以及姿态的局部空间变形的不变特性。同时,为了解决与运动模板的时间对齐问题,他们采用动态时间规整(Dynamic Temporal Warping,DTW)的模板匹配方法来进行动作识别。虽然该方法对于无噪声的 MoCap 数据具有很好的扩展性和有效性,但是对于具有高复杂度的人体动作,例如“跳舞”,要定义出具有强判别性和鲁棒性的关节点间的几何关系并非易事。

即使是同类动作序列,MoCap 数据也存在很大的时-空差异。空间上的差异主要来自于动作执行过程中对象的位置、角度以及相机视角等的变化,其可以用一个刚体变换矩阵来刻画;而时间上的非线性差异则主要是由于不同对象的执行过程差异,或者是同一对象的 2 次不同节奏的执行。因此,需要有一个合适的空间表示和时间对齐方法来进行动作的比较。基于此,文献[13]提出了一种新的基于距离矩阵的不变特征来表示运动姿态。该距离矩阵的每个元素定义为任意两个关节点之间的欧式距离。为了克服对象的个体外表差异,在计算距离之前进行了归一化处理。由于刚体变换具有距离保持不变性,因此,具有相等关节点的距离矩阵所描述的动作姿态相同。从而使得该特征具有鲁棒的空间不变性。最后,文中采用动作图(Action Graph)来描述运动的动态性,而转移模型通过学习获得,从而有效地解决了时间差异,即时间对齐问题。

1.1.2 基于深度相机的关节点方法

深度相机能够感知距离信息,并通过 OpenNI 算法[14]来实时估计人体的骨架关节点位置。但这些关节点数据含有较大的噪声,因此,很难将基于 MoCap

数据的方法直接迁移过来。近来,学术界提出了不少有效的针对深度相机的关节点方法,主要是采用关节的位置及其相对位置关系来建模人体姿态。

1) 关节点位置

基于关节点位置的方法多采用词袋(Bag of Word,BoW)技术来建模人体运动姿态。文献[15]首先用关节位置处的时-空描述子来构造视觉单词,每个视觉单词表示一个具有强判别性的运动姿态。然后将动作实例用词典中单词出现的频率来表示,从而获得每个动作的一个视觉单词直方图,并作为 SVM 的输入来对动作进行分类。Xia 等人[16]提出了一种基于关节点位置直方图(Histograms of 3D Joint Locations, HOJ3D))的紧凑人体姿态表示方法。他们首先采用修正的球坐标系统将 3D 空间划分 84 个特定方向,每个方向作为直方图的一个 bin。然后将关节点位置投影到这 84 个方向 bins 上构成 HOJ3D,并采用 LDA 对特征空间进行降维以获得更加鲁棒的特征表示。随后,将降维后的特征空间聚类为 k 个姿态视觉单词并进行量化,每个视觉单词表示动作的一个原型姿态。最后,采用离散隐马尔科夫模型来建模视觉单词的时间进化过程并进行分类。

2) 关节点间的时-空相对位置关系

关节点间的时-空相对位置关系能够提供更丰富的动作姿态与运动信息。文献[17-18]提出了一种新的名为 EigenJoints 的特征,它是姿态特征 f_{cc}、运动特征 f_{cp}和偏移特征 f_{ci}三个子特征的组合。以单帧图像内关节点之间的位置差作为静态的姿态特征,而运动特征则由关节点在前后相邻两帧间的位置差构成。偏移特征表示为当前帧与初始帧内对应关节点的位置差,前提是假设初始姿态为中立姿态。然后对这三个通道的特征进行归一化和 PCA 降维处理,以获得低冗余度和低噪声的 EigenJoints 描述子。最后采用朴素 Bayes 最近邻(Naive Bayes Nearest Neighbor,NBNN)分类器进行分类。与此类似,文献[19]同样是采用 NBNN 分类器,所不同的是其采用关节点位置在一定时间段内的偏移作为动作的描述特征,并结合视觉词袋模型来进行动作识别。

3) 低延时(Low Latency)的动作识别

最近有一些研究者将工作重点转移到处理更具挑战性的在线动作识别系统或交互式动作识别问题上,即低延时(Low Latency)的动作识别。所谓延时,指的是开始执行动作的时间与该动作被正确识别出来的时间之差。该类问题的主要挑战在于识别精度与低延时之间要有一个合理的权衡。为此,Ellis 等人[20]提出了一种基于 3D 关节点位置序列的延时感知(Aware)学习方法来渐少动作识别时的延时。他们使用一个具有延时感知的多实例学习算法来训练一个基于逻辑回归的分类器用以自动从 3D 关节点数据序列中抽取最具强判别性的典型人体姿态并实时用这些姿态与模板库中的人体姿态进行比对以尽可能早地对动作类别进行判决。文献[21]提出一种快速、简单且强有力的运动姿态(Moving Pose,MP)特征来进行低延时的人体动作识别。MP 描述子由关节点的位置、速度和加速度信息 3 个元素构成。前者用于描述动作的静态姿态特征,而后两者对运动的动态过程进行刻画。最后采用改进的 KNN 分类器来进行识别。

4) 流形学习方法

上述方法均在欧式空间中进行人体动作识别,然而,新近出现了一些基于非欧式空间的流形学习方法。Devanne 等人[22]提出了一种人体动作的紧凑表示方法,将关节点的 3D 坐标及其随时间的变化表示为动作空间中的一条轨迹,从而将动作识别归结为在 Riemannian 流形上计算轨迹形状的相似性,并采用 KNN 分类器进行动作的分类。该表示能够同时捕获人体动作的形状和动态性,这些特征对于建模人体动作来说至关重要。与此类似,文献[23]采用自回归移动平均模型(Auto Regressive and Moving Average model,ARMA)来建模人体关节点运动轨迹的动态过程,最终将动作的分类问题转化为 ARMA 观测矩阵的比较问题。ARMA 观测矩阵的列向量所张成的子空间为 Grassmann 流形上的一个点,该点可用局部切丛(Local Tangent Bandle,LTB)来表示。最后将 LTB 作为 SVM 分类器的输入来进行动作识别。

1.2 基于深度映射图像的方法

相对于精简的骨架关节点数据而言,深度映射图像能够提供更丰富的有关动作的形状和运动信息。基于深度映射图像的人体动作识别技术可分为整体方法和局部方法 2 种。下面分别对其进行介绍。

1.2.1 整体方法

整体方法通常使用投影和时-空体这样的全局特征来建模人体动作姿态和运动的动态性。

1) 投影方法

投影方法首先将深度映射序列所构成的四维时-空体投影到特定平面视图中,然后提取有意义的描述子来建模人体动作。Yang 等人[24]提出了一种称为深度运动映射(Depth Motion Map,DMM)的方法来捕获时间聚集的运动能量。具体来说,首先

将深度映射图像投影到前视图、顶视图和侧视图3正交平面上并归一化。接着,在每个投影视图下计算前后连续2帧图像的差分并阈值化后产生一个二值映射图像,并将二值映射图像累积叠加起来以获得每个投影视图的DMM。然后,方向梯度直方图(Histogram of Oriented Gradients, HOG)被应用到每个视图来提取特征,最后将3个视图的特征串接起来形成DMM-HOG描述子,并用SVM分类器进行识别。同样,Chen等人[25]提出的TriViews框架也是基于前、顶和侧视图投影,分别在3个投影视图中提取轨迹形状、运动边界和时-空兴趣点特征,并以概率融合方法(Probabilistic Fusion Approach, PFA)[26]来进行多种特征的有效组合,从而形成一种基于3视图的通用动作识别框架。

2) 时-空体方法

与投影方法不同,文献[27]直接将深度映射序列作为一个时-空四维体,然后以不同位置和尺寸随机采样该四维体来获得子体并计算子体中的占有信息(即落入子体中的像素个数),以此构成一种新的随机占有模式(Random Occupancy Pattern,ROP)特征,随后对这些特征进行稀疏编码,其编码系数作为SVM的输入来进行人体动作识别。与此类似,Vieira等人[28]提出的时-空占有模式(Space-Time Occupancy Pattern, STOP)特征也对子体中的占有信息进行编码,所不同的是,他们采用相同尺寸的网格来采样四维体。

1.2.2 局部方法

相对于整体方法,基于深度映射图像序列的局部特征方法首先提取兴趣点,然后在兴趣点周围邻域内计算一个局部特征描述子来进行动作识别。较整体方法而言,局部方法对噪声、视角和遮挡具有更好的鲁棒性。由于采集的深度序列图像中混有噪声,将RGB图像中广泛使用的兴趣点检测子,如Cuboid、3D Hessian和3D Harris角点,简单地拓展到深度图像中效果并不理想。近来,出现了一些针对深度数据的局部特征描述子。Cheng等人[29]提出一种新的深度信息描述子:比较编码描述子(Comparative Coding Descriptor),用于刻画动作体中时-空点的结构关系。他们首先使用3D Harris角点检测算法提取时-空兴趣点作为显著点,然后以显著点为中心构造一个$3\times3\times3$的Cuboid,在此Cuboid中比较中心点与其它26个点的深度值,将差值按顺序进行编码从而构成CCD特征。文献[30]构建了一种局部深度模式(Local Depth Pattern,LDP),用它来描述以特征点为中心的局部邻域内像素的深度值差异。最近,Xia等人[31]提出以DCSF(Depth Cuboid Similarity Feature)作为时-空深度Cuboid的描述子。DCSF主要是基于内容的自相似性来描述深度视频的局部外观模式,即通过施加共享相同的视觉模式约束来捕获视频局部Patch的内部几何分布信息。该方法同样具有多模态适应性,能够同时适用于RGB和深度数据。

1.3 基于深度相机3D点云数据的方法

通过深度相机采集的深度映射图像可以直接转换为3D点云数据。利用点云数据的动作识别可分为基于局部占有模式和基于曲面法线2种方法。前者主要是以关节点为锚点,计算其周围邻域内点云数据的局部占有模式(Local Occupancy Pattern, LOP)特征[32],主要用于描述动作的类内变异。而基于法线的方法则在3D点云序列中的每个点的局部时-空体邻域内提取曲面的法线特征来描述动作的几何和运动信息[33-35]。目前,这方面的研究相对较新颖,相关的文献也较少。

1.3.1 局部占有模式

人与周围物体和环境的交互时有发生,使得采集的深度图像有可能产生严重的遮挡,最终导致跟踪的关节点精度急剧下降甚至完全错误,从而增加了动作的类内可变性。为此,文献[32]提出了一种新的LOP特征。它以关节点周围区域点云数据为基础,首先将该局部空间划分为一定大小的网格,然后计算每个网格内的占有信息,并通过Sigmoid函数归一化后构成LOP特征向量。除此之外,他们还将关节点特征与LOP特征串接起来进行短时傅立叶变换,并将其系数作为每个关节点的傅里叶时间金字塔特征。这样生成的特征无需严格的时间对齐,并对噪声具有很好的鲁棒性。为了获得对每个动作有判别性的关节点组合,即Actionlet,他们引入了基于先验的数据挖掘方法并通过多核学习算法来获得该动作的Actionlet集成模型。

1.3.2 曲面法线

与基于局部占有模式的方法不同,Oreifej等人[34]采用直方图来捕获点云序列所构成的4D曲面法线的方向分布(Histogram of Oriented 4D surface Normals, HON4D)。为了构建HON4D,首先将点云序列看作一个由时间、深度和空间坐标构成的4D空间,并用600个单元、120个顶点的四维体对该4D空间进行初始量化,每个顶点称为一个投影子(Projector)。接着以差分的方式计算该4D空间的曲面法线方向并投影到这120个投影子上。由于计算HON4D描述子需要在整个视频序列上进行,不

能够进行实时动作识别,且该方法首先需要对动作进行时-空对齐,因此具有一定的局限性。与HON4D特征类似,最近,Yang等人[35]则通过聚类点云序列中每个点邻域内的4D法线形成新的超级法向量描述子(Super Normal Vector, SNV)来扩展HON4D,它能够同时捕获局部运动和几何信息,因此具有更强的描述能力。文献[33]提出了一种新的点云数据描述子:方向主成分直方图(Histogram of Oriented Principal Components, HOPC)。该方法以3D点云序列中的每个点 P 为中心,计算落入半径为 r 的球体内所有点的散布矩阵(Scatter Matrix)的3个特征向量,即3个主成分向量,并将它们投影到以正20面体的每个顶点向量所确定的方向上。最后将投影分量串接起来,构成一个大小为60维的描述向量作为该点的HOPC描述子,用于从局部几何形状上来刻画动作。该方法本质上仍然是对以3D点云所确立的曲面法线进行方向直方图编码,因为局部曲面的最小主成分是该曲面法线的最小二乘估计。因此,较基于差分的方法[34-35],HOPC对噪声具有更好的鲁棒性。

1.4 RGB图像与深度信息融合的多模态(Multi-Modal)方法

微软Kinect相机的输出是一个多模态信号,它能够同时提供RGB视频、深度映射图像序列以及骨架关节点信息,从而有效克服了由于传统的RGB相机将3D物理世界投影到2D图像平面而失去深度信息并导致物体间的空间位置关系丢失、运动目标检测与上下环境建模的精度显著下降的问题。同时,不同模态下的特征彼此相互独立但又具有很强的互补性。为了获得更好的识别性能,目前一些研究工作将多模态下的特征进行有效融合并设计出了具有高判别性的描述向量,即同时利用深度、骨架与视觉信息来改进识别结果。其融合方式可分为特征层和决策层2种融合模式。

1.4.1 特征层融合

特征层融合又称为早期融合,是一种较简单、直接的融合方法,即将不同模态下提取的特征向量组合起来以形成一个更具描述力的向量。文献[36]提出了一种将深度数据的骨架特征与RGB图像中提取的基于剪影的特征进行直接串接融合的方法。骨架特征按文献[37]所描述的方法进行了归一化,使其具有尺度与旋转不变性;而基于剪影的特征则通过径向提取剪影的轮廓点来生成[38]。该融合方法能够同时保留由两种模态下的特征所提供的特性信息,融合后的特征仍然是一个相对低维的向量,适合实时识别。Shahroudy等人[39]提出了一种无监督的结构化稀疏特征融合方法。他们分别提取了RGB通道中的密集轨迹特征与深度通道中的骨架数据,然后对这些特征进行加权组合,权重为无监督的结构化稀疏表示系数。该方法有效避免了由于不同模态下的特征直接进行组合时有可能降低其判别性能的问题。

1.4.2 决策层融合

不同于特征层的融合,决策层的融合又称为后期融合,主要是将针对各模态特征分别设计的分类器输出结果进行组合以作最终的判决。文献[40]分别从RGB图像中提取3D Harris角点、时-空梯度自相关特征[41]和从深度数据中提取Eigen-Joints[17-18]、HON4D[34]特征进行决策层的融合,并详细讨论了5种不同的决策层融合方法对识别性能的影响,具体流程如图1所示。

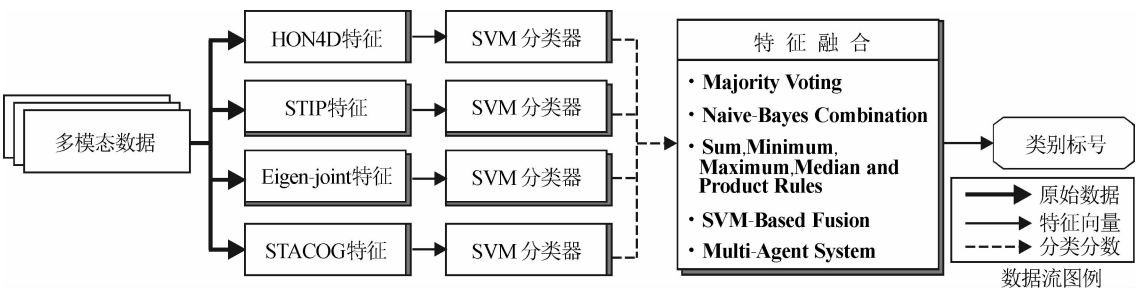

图1 文献[40]提出的决策层融合方案

Fig. 1 The schemes of the decision-level/late fusion proposed by [40]

1.5 基于深度学习(Deep Learning)的方法

深度学习是一种无监督的学习方法,相对于人工设计的特征,它能够从大量训练样本数据中自主学习到具有强判别力和表示能力的特征。随着深度学习在语音识别、自然语言处理以及计算机视觉领域的成功应用,人体动作识别研究领域也相继出现了一些基于该技术的方法。

Cho Kyunghyun等人[42]提出了一种基于深度

神经网络(Deep Neural Network,DNN)的人体动作识别方法。他们首先从每帧中抽取以下特征:关节点的相对位置、时间差分和归一化的运动轨迹,然后基于这些特征来训练一个多层感知机(Multi Layer Perceptron,MLP)并用于对输入数据进行重构和分类。

最近,Wang 等人[43]设计了一种多层深度运动映射的 CNN 架构来进行人体动作识别。首先,为了有效提取与视角无关的人体形状与运动信息,他们对原始 3D 点云数据进行了旋转并投影到前视图、顶视图和侧视图 3 正交平面视图中,然后对每个视图在不同时间尺度上生成多层深度运动映射图(Hierarchical Depth Motion Maps,HDMM)并输入到 CNN 网络中进行特征提取,最后对 3 个视图通道的识别结果进行决策层的融合。

1.6 对各方法的总结

在此节中,笔者对综述文献分别从文献所属的分类类别、发表的时间、文中所使用的特征及对特征的表示方法、所采用的分类器和实验中用到的测试数据集 5 个指标进行了总结,如表 1 所示。

表 1 按照本文提出的分类方法对综述文献的汇总

Tab. 1 Summary of methods for action recognition based on our proposed taxonomy

分类法		文献	发表时间	特征及表示	分类器	数据集
骨架关节点	MoCap	[44]	2009	关系运动特征	DTW 模板匹配	CMU MoCap[11], HDM05[45]
		[13]	2012	距离矩阵	动作图	HDM05[45]
	OpenNI	[15]	2010	关键点位置+BoW	SVM	MSR Action3D[15]
		[16]	2012	HOJ3D	离散隐马尔科夫模型	MSR Action3D[15], UTKinect Action[16]
		[17] [18]	2012 2014	EigenJoints	NBNN	MSR Action3D[15]
		[19]	2015	关节点的时-空偏移+BoW	NBNN	UCF Kinect[20], MSRC-12 KinectGesture[46]
		[20]	2013	延时感知特征	逻辑回归分类器	UCF Kinect[20], MSRC-12 KinectGesture[46], MSR Action3D[15]
		[21]	2013	MP 特征	KNN	MSR Action3D[15], MSR DailyActivities3D[32]
		[22]	2014	关节点运动轨迹+流形学习	KNN	MSR Action3D[15], UTKinect Action[16], UCF Kinect[20], Florence 3D Action[47]
		[23]	2015	ARMA+流形学习	SVM	MSR Action3D[15], UTKinect Action[16], UCF Kinect[20]
深度映射图像	整体方法	[24]	2012	DMM+HOG	SVM	MSR Action3D[15]
		[25]	2015	轨迹形状+运动边界+时-空兴趣点特征	SVM+ PFA 决策层融合	MSR Action3D[15], UTKinect Action[16], MSR DailyActivities3D[32]
		[27]	2012	ROP	SVM	MSR Action3D[15], MSR Gesture3D[48]
		[28]	2012	STOP	动作图	MSR Action3D[15]
	局部方法	[29]	2012	CCD	SVM	ACT4 Dataset[29]
		[30]	2012	LDP	SVM	RGBD-HuDaAct[49]
		[31]	2013	DCSF+PCA	SVM	MSR Action3D[15], MSR DailyActivities3D[32]
3D 点云	局部占有模式	[32]	2012	LOP	SVM	MSR Action3D[15], MSR DailyActivities3D[32], CMU MoCap[11]
		[33]	2014	HOPC	SVM	MSR Action3D[15], MSR Gesture3D[48], MSR ActionPairs[34], UWA 3D Multiview Activity[33]
	曲面法线	[34]	2013	HON4D	SVM	MSR Action3D[15], MSR Gesture3D[48], MSR DailyActivities3D[32], MSR ActionPairs[34]
		[35]	2014	SNV	SVM	MSR Action3D[15], MSR Gesture3D[48], MSR DailyActivities3D[32], MSR ActionPairs[34]
RGB+深度信息	特征层融合	[36]	2013	骨架特征+剪影特征	DTW	MSR Action3D[15]
		[39]	2014	RGB 通道密集轨迹特征+骨架数据	SVM	MSR DailyActivities3D[32]
	决策层融合	[40]	2015	3D Harris 角点+时-空梯度自相关特征+EigenJoints+HON4D	SVM	MSR Action3D[15], MSR DailyActivities3D[32], UTKinect Action[16], CAD-60[50]
深度学习		[42]	2013	关节点的相对位置+时间差分+归一化的运动轨迹	MLP	HDM05 MoCap[45]
		[43]	2015	HDMM	CNN	MSR Action3D[15], MSR DailyActivities3D[32], UTKinect Action[16]

各种方法所采用的数据集和验证策略差别很大,因此很难采用一个统一标准来对它们的性能优劣做出一个客观、真实的评价。由于 MSR Action3D[15]数据集是最早建立的、使用最广泛的基于深度相机的公用测试集,大多数方法都将其作为性能评价的一个基准。但尽管这样,仍然有许多方法在测试样本与训练样本的划分方式和数量上存在很大差异。

因此,笔者以 MSR Action3D 数据集为讨论基础,采用跨目标(Cross-subject)和交叉(Cross-validation)验证两种验证方式以从不同角度来全面汇总和比较各种算法的性能差异。

首先对 MSR Action3D 数据集做一个简单介绍。MSR Action3D 数据集包含 20 个类别的动作数据,由 10 个演员执行,每类动作每个演员执行 2~3次,总共构成 567 个样本序列。该数据集的其他描述信息见表 2。为了减少测试时的计算复杂度,数据集依据动作的复杂程度被再次划分为 3 个子集:AS_1,AS_2 和 AS_3。其中子集 AS_1 和 AS_2 中包含的动作复杂度相对较低,但每个子集内的动作相似度较高。子集 AS_3 中的动作复杂度最高。每个子集具体包含的动作类别如表 3 所示。

表 2 对综述文献中所引用数据集的汇总

Tab. 2 Summary of datasets for action recognition cited in the reviewed literature

数据集	时间	类别数	样本数	特性	使用该数据集的文献
CMU MoCap[11]	2003	144 个大类	2 605	目前规模最大的 MoCap 数据集;没有按照特定的脚本执行动作,而是随机采样执行的动作,类内、类间差异巨大;提供低分辨率的 RGB 视频和 3 种格式的关节点数据:tvd,c3d,amc	[32],[44]
HDM05 MoCap[45]	2007	130	2 601	5 个演员,每类动作每个演员执行若干次;类内差异大,每个动作执行实例有 10~50 次	[13], [42], [44]
MSR Action3D[15]	2010	20	567	10 个演员,每类动作每个演员执行 2~3 次;提供 20 个关节点的 3D 坐标数据、深度图像与 RGB 图像;视频序列为无背景的纯人体运动目标	[15-18], [20-25], [27-28], [31-36], [40], [43]
UTKinect Action[16]	2012	10	200	10 个演员,每类动作每个演员执行 2 次;提供 20 个关节点的 3D 坐标数据	[16], [22-23], [25], [40], [43]
UCF Kinect[20]	2013	16	1 280	16 个演员,每类动作每个演员执行 5 次;提供 15 个关节点的 3D 坐标及方向数据	[19-20], [22-23]
MSRC-12 KinectGesture[46]	2012	12	594	30 个演员,每类动作每个演员执行 3 次;提供 20 个关节点的 3D 坐标数据	[19-20]
MSR DailyActivities3D[32]	2012	16	320	10 个演员;大部分样本涉及到人与物体的交互;捕获的 3D 关节点坐标受噪声污染严重	[21], [25], [31-32], [34-35], [39-40], [43]
Florence 3D Action[47]	2012	9	215	10 个演员,每类动作每个演员执行 3 次;动作相似性大,包含人与物体的交互,同类动作具有不同的执行方式	[22]
MSR Gesture3D[48]	2012	12	336	10 个演员,每类动作每个演员执行 2-3 次;提供深度图像与 RGB 图像,样本存在相当普遍的自遮挡	[27], [33-35]
ACT4 Dataset[29]	2012	12	6 844	24 个演员,由 4 个相机从不同视角采集日常生活活动视频;提供深度图像与 RGB 图像	[29]
RGBD-HuDaAct[49]	2011	12	1 189	30 个演员,每类动作每个演员执行 2~4 次;提供深度图像与 RGB 图像,样本中混有随机背景动作	[30]
MSR ActionPairs[34]	2013	6	180	10 个演员,每类动作每个演员执行 3 次;每个动作对有相似的运动和形状	[33-35]
UWA3D Multiview Activity[33]	2013	30	720	10 个演员,每类动作每个演员执行 2~3 次;存在自遮挡和高度相似性;具有视角和尺度变化;提供关节点的 3D 坐标数据、深度图像、深度的前景分割图像与 RGB 图像	[33]
CAD-60[50]	2012	12	60	4 个演员,在 5 个不同的场景中执行动作;提供 15 个关节点的 3D 坐标数据、深度图像与 RGB 图像	[40]

表 3 MSR Action3D 数据集动作子集的划分

Tab. 3 Actions in each of the MSR Action3D subsets

AS_1		AS_2		AS_3	
标签	动作类型	标签	动作类型	标签	动作类型
a02	Horizontal arm wave	a01	High arm wave	a06	High throw
a03	Hammer	a04	Hand catch	a14	Forward kick
a05	Forward punch	a07	Draw cross	a15	Side-kick
a06	High throw	a08	Draw tick	a16	Jogging
a10	Hand clap	a09	Draw circle	a17	Tennis swing
a13	Bend	a11	Two-hand wave	a18	Tennis serve
a18	Tennis serve	a14	Forward kick	a19	Golf swing
a20	Pick-up and throw	a12	Side-boxing	a20	Pick-up and throw

1.6.1 跨目标验证(Cross-subject test)

跨目标测试的思想是:训练样本与测试样本分别来自不同演员执行的动作序列。即使是同类型的动作,由于个体在执行时的差异性,往往使得采集的样本具有较大的类内方差。该类验证机制可以有效评估算法的泛化性能和鲁棒性。在此数据集上,有 9 篇文献采用以下策略来进行测试:对于每个子集中的样本数据,由 1、3、5、7、9 号演员执行动作所采集的样本作为训练数据,剩余样本作为测试数据。按照方法、时间、每类子集上的识别率(%)、平均识别率(%)和在整个数据集上的识别率(%)进行汇总与比较,具体内容如表 4 所示。另外,有 3 篇文献采用 1～5 号演员执行动作所采集的样本作为训练数据,剩余样本作为测试数据。其识别率如表 5 所示。

表 4 跨目标测试性能比较

Tab. 4 Cross-subject test results

方法	年份	各子集上的识别率/%			平均识别率/%	整体识别率/%
		AS_1	AS_2	AS_3		
Action recognition based on a bag of 3d points, Li, et al.[15]	2010	72.90	71.90	79.2	74.67	—
View invariant human action recognition using histograms of 3D joints, Xia, et al.[16]	2012	87.98	85.48	63.46	78.97	—
Effective 3D action recognition using EigenJoints, Yang and Tian[18]	2014	—	—	—	83.30	—
Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition, Ellis, et al.[51]	2013	—	—	—	65.70	—
Robust 3D Action Recognition with Random Occupancy Patterns, Wang, et al.[27]	2012	—	—	—	—	86.50
STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences, Vieira, et al.[28]	2012	84.70	81.30	88.40	84.80	—
Mining actionlet ensemble for action recognition with depth cameras, Wang, et al.[32]	2012	—	—	—	—	88.20
Super Normal Vector for Activity Recognition Using Depth Sequences, Yang and Tian[35]	2014	—	—	—	93.09	—
Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices, Chaaraoui, et al.[36]	2013	92.38	86.61	96.40	91.80	—

说明:1、3、5、7、9 号演员作为训练样本,其余作为测试样本。

表 5 跨目标测试性能比较

Tab. 5 Cross-subject test results

方法	年份	各子集上的识别率/%			平均识别率/%	整体识别率/%
		AS_1	AS_2	AS_3		
The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection, Zanfir, et al.[21]	2013	—	—	—	91.70	—
HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, Oreifej and Liu[34]	2013	—	—	—	—	88.89
Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera, Xia and Aggarwal[31]	2013	—	—	—	89.30	—

说明:1～5 号演员作为训练样本,其余作为测试样本。

1.6.2 交叉验证(Cross-validation)

交叉验证是用来验证分类器性能的一种常用统计分析方法,基本思想是按照一定的划分方式将原始数据集进行分组,一部分作为训练集,另一部分作为验证集。首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型,以此来作为评价分类器的性能指标。根据其划分方式不同可分为:2-折交叉验证(2-fold cross-validation)和留一法交叉验证(Leave-One-Out cross-validation)等。前者随机将原始数据均分为 2 组,将每组分别做一次验证子集和训练子集,这样会得到 2 个模型,用在 2 个模型最终的验证子集的分类精度的平均值作为分类器的性能指标。文献[37]便采用此种验证方式,其识别性能如表 6 所示。

而对于留一法交叉验证,其基本思路是:假设原始数据有 N 个样本,每个样本单独作为验证集,其余的 $N-1$ 个样本作为训练集,这样会得到 N 个模型,用这 N 个模型最终的验证集的分类准确率的平均数作为分类器的性能指标。留一法评估结果的客观度和准确性较前面几种验证方法均要高,但计算量也是该方法的致命弱点。文献[36]采用此种验证方法,其结果如表 6 所示。

表 6 交叉验证

Tab. 6 Cross-validation test

交叉验证方法	方法	年份	各子集上的识别率/%			平均识别率/%	整体识别率/%
			AS_1	AS_2	AS_3		
2-折交叉验证	Evolutionary joint selection to improve human action recognition with RGB-D devices, Chaaraoui, et al.[37]	2014	91.59	90.83	97.28	93.23	—
留一法交叉验证	Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices, Chaaraoui, et al.[36]	2013	90.65	85.15	95.93	90.58	—

2 公用测试数据集

人体动作识别技术的巨大进步得益于各种公用标准测试数据集的建立。笔者对综述文献中引用的 14 个数据集从建立的时间、所包含的动作类别数、样本数和特性等角度进行了总结,如表 2 所示。这些数据集中的绝大多数均采用微软的 Kinect 相机作为采集工具,它们为各种动作识别算法的性能分析搭建了一个公平的环境,并将继续推动和促进相关研究工作的进一步发展。

3 未来的研究方向

低成本的深度传感器为人体动作识别的研究提供了极好的发展机会,在过去的几年中,基于深度相机的人体动作识别研究虽然取得了巨大的进步,但仍然存在许多挑战亟待解决。在未来的研究工作中,一方面要从深度和骨架数据中设计更具判别力和紧致的特征向量来描述人体动作,另一方面是拓展当前已有的方法来应对更加复杂的人体动作,如交互和群体活动等,具体来说,将涉及到以下 4 个方面的问题。

1) 交互动作与群体活动识别

理解与解释人-人、人-物之间的交互动作/活动的语义信息在诸如机器人和人-机交互等领域中有着广泛而实际的应用。交互动作中必然会产生身体部分被遮挡、人与物间的相互遮挡和复杂背景问题,从而使得各种动作描述方法的刻画、判别能力显著下降。目前的动作识别方法主要聚焦在手势与简单的动作识别上,仅有少数工作针对较为复杂的人-人或人-物交互动作的识别。在这方面仍需要深入的研究来取得更好的处理能力以满足实际应用的需要。另外,在包括多个人体目标的群体活动识别方面,目前还没有相关的研究工作,其中一个主要原因是当前的低成本深度传感器还不具备捕获大范围场景数据的能力。

2) 多视角与跨视角动作识别

视角变化问题,无论是在传统的 RGB 图像中,还是在深度数据中,对于动作识别来说始终是一个大的挑战。在现实世界的动作识别系统中,视频序列往往是以任意相机视点捕获;因此,系统性能必须具备视角不变性。然而,目前的大多数算法仍然是基于有约束的视角,如要求演员正对着相机执行动作。此外,当前被广泛使用的标准测试数据集绝大部分也是在单一视角下拍摄的。由于视角的随意性、动作的多变性以及训练样本的有限性,实际应用中更一般的情况是在某一视角下的特征仍然适用于其它视角下的识别任务,即交叉视角下的动作识别问题。因此,未来需要更复杂的视角不变算法来应

对这些问题。而从深度相机捕获的点云数据中设计与提取视角不变特征将是一种行之有效的方法。

3) 低延时动作识别

目前的人体动作识别方法在追求高识别率的同时其算法的复杂度也随之增加,即具有高的计算延时性。而作为动作识别的一个重要应用领域,人-机交互对算法的实时性要求非常高,甚至在一个动作还没有完全执行完整时就能准确判断或预测动作所传达的语义信息,称之为低观测延时性。因此,设计与开发具有低延时性的动作识别算法是朝着未来实用型人-机交互系统迈出的关键一步。

4) 多模态特征与融合技术

不同模态下的人体动作数据具有各自的特性,同时拥有很强的描述互补性,能够进一步提高动作识别系统的性能,这一点已经在目前一些基于多模态特征融合的方法中得到了强有力的证实。充分利用和挖掘深度传感器所提供的多模态数据并进行强有效的融合仍然是未来设计高精度人体动作识别系统的一个行之有效的途径。当前,仍然没有一种具有绝对优势的特征描述算法来进行有效的人体动作识别,能否将不同模态下的各种动作识别方法统一在一个框架下进行,即设计出具有适用于不同模态数据的自适应特征检测与描述算法将是另一个新的起点。

总之,人体动作识别系统不仅需要从视觉信号中提取低层的外观及运动信息,而且还需要复杂的机器学习模型来理解这些信息的语义含义。为此,需要综合利用多方面的知识,包括采用兼具视觉与深度信息的多模态传感器来精确地获得视觉信号,能有效描述高维视觉数据的视频或图像表示方法,发掘有意义知识的模式挖掘技术以及能够从大数据中进行自主学习的机器学习算法。

4 结 语

深度传感器的发展,尤其是微软 Kinect 相机的出现,为人体动作识别提供了新的机遇。越来越多的方法采用深度数据来进行人体动作识别。本文对最近几年中基于深度数据的人体动作识别算法进行了较全面的综述。并对相关的标准测试数据集进行了较详细的描述。最后给出了未来的研究方向。

参考文献:

[1] Ramanathan M, Yau Wei-Yun, Teoh Eam Khwang. Human action recognition with video data: research and evaluation challenges[J]. Human-Machine Systems, IEEE Transactions on, 2014, 44(5): 650-663.

[2] Weinland Daniel, Ronfard Remi, Boyer Edmond. A survey of vision-based methods for action representation, segmentation and recognition[J]. Computer Vision and Image Understanding, 2011, 115(2): 224-241.

[3] Weinland D, Boyer E. Action recognition using exemplar-based embedding[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008:1-7.

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

[5] Guo K, Ishwar P, Konrad J. Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels[C]//Recognizing patterns in signals, speech, images, and videos, 2010: 294-305.

[6] Gorelick L, Blank M, Shechtman E, et al. Actions as space-time shapes[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007, 29(12): 2247-2253.

[7] Klaser Alexander, Marszalek Marcin. A spatio-temporal descriptor based on 3d-gradients[C]. British Mach. Vision Conf., 2008.

[8] Wang Heng, A Klaser, C Schmid, et al. Action recognition by dense trajectories[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011: 3169-3176.

[9] Han Jungong, Shao Ling, Xu Dong, et al. Enhanced computer vision with microsoft kinect sensor: A review[J]. Cybernetics, IEEE Transactions on, 2013, 43(5): 1318-1334.

[10] Johansson Gunnar. Visual motion perception[J]. Scientific American, 1975, 232(6): 76-88.

[11] Cmu: Carnegie-mellon mocap database[EB/OL]. [2014-12-21]. http://mocap.cs.cmu.edu/, 2003.

[12] Müller Meinard, Röder Tido, Clausen Michael. Efficient content-based retrieval of motion capture data[C]//ACM Transactions on Graphics (TOG), 2005: 677-685.

[13] Vieira A W, Lewiner T, Schwartz W R, et al. Distance matrices as invariant features for classifying mocap data[C]//Pattern Recognition (ICPR), 2012 21st International Conference on, 2012: 2934-2937.

[14] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011: 1297-1304.

[15] Li Wanqing, Zhang Zhengyou, Liu Zicheng. Action

recognition based on a bag of 3d points[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010: 9-14.

[16] Xia Lu, Chen Chia-Chih, Aggarwal J K. View invariant human action recognition using histograms of 3d joints[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, 2012: 20-27.

[17] Yang Xiaodong, Tian Yingli. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor [C]// Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, 2012: 14-19.

[18] Yang Xiaodong, Tian Yingli. Effective 3d action recognition using eigenjoints[J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 2-11.

[19] Lu Guoliang, Zhou Yiqi, Li Xueyong, et al. Efficient action recognition via local position offset of 3d skeletal body joints[J]. Multimedia Tools and Applications, 2015: 1-16(DOI: 10.1007/s11042-015-2448-1).

[20] Ellis Chris, Masood SyedZain, Tappen MarshallF, et al. Exploring the trade-off between accuracy and observational latency in action recognition[J]. International Journal of Computer Vision, 2013, 101(3): 420-436.

[21] Zanfir M, Leordeanu M, Sminchisescu C. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection[C]//Computer Vision (ICCV), 2013 IEEE International Conference on, 2013: 2752-2759.

[22] Devanne M, Wannous H, Berretti S, et al. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold[J]. Cybernetics, IEEE Transactions on, 2014, PP(99): 1-1.

[23] Slama Rim, Wannous Hazem, Daoudi Mohamed, et al. Accurate 3d action recognition using learning on the grassmann manifold[J]. Pattern Recognition, 2015, 48(2): 556-567.

[24] Yang Xiaodong, Zhang Chenyang, Tian Yingli. Recognizing actions using depth motion maps-based histograms of oriented gradients[C]//Proceedings of the 20th ACM international conference on Multimedia, 2012: 1057-1060.

[25] Chen Wenbin, Guo Guodong. Triviews: A general framework to use 3d depth data effectively for action recognition[J]. Journal of Visual Communication and Image Representation, 2015, 26: 182-191.

[26] Guo Guodong, Fu Yun, Dyer Charles R, et al. A probabilistic fusion approach to human age prediction [C]//Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, 2008: 1-6.

[27] Wang Jiang, Liu Zicheng, Chorowski Jan, et al. Robust 3d action recognition with random occupancy patterns[C]//Computer vision-eccv 2012, Springer Berlin Heidelberg, 2012: 872-885.

[28] Vieira Antonio W, Nascimento Erickson R, Oliveira Gabriel L, et al. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. [C]//Progress in pattern recognition, image analysis, computer vision, and applications, Springer Berlin Heidelberg, 2012: 252-259.

[29] Cheng Zhongwei, Qin Lei, Ye Yituo, et al. Human daily action analysis with multi-view and color-depth data[C]//Computer Vision-ECCV 2012. Workshops and Demonstrations, 2012: 52-61.

[30] Zhao Yang, Liu Zicheng, Yang Lu, et al. Combing rgb and depth map features for human activity recognition [C]//Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, 2012: 1-4.

[31] Lu Xia, Aggarwal J K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013: 2834-2841.

[32] Wang Jiang, Liu Zicheng, Wu Ying, et al. Mining actionlet ensemble for action recognition with depth cameras[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012: 1290-1297.

[33] Rahmani Hossein, Mahmood Arif, Huynh Du Q, et al. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition[C]//Computer vision-eccv 2014, Springer, 2014: 742-757.

[34] Oreifej Omar, Liu Zicheng. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013: 716-723.

[35] Yang Xiaodong, Tian Yingli. Super normal vector for activity recognition using depth sequences[C]//Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014: 804-811.

[36] Chaaraoui Alexandros Andre, Padilla-López José Ramón, Flórez-Revuelta Francisco. Fusion of skeletal

and silhouette-based features for human action recognition with rgb-d devices[C]//Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, 2013: 91-97.

[37] Chaaraoui Alexandros Andre, Padilla-López José Ramón, Climent-Pérez Pau, et al. Evolutionary joint selection to improve human action recognition with rgb-d devices [J]. Expert Systems with Applications, 2014, 41(3): 786-794.

[38] Chaaraoui Alexandros Andre, Flórez-Revuelta Francisco. Human action recognition optimization based on evolutionary feature subset selection[C]//Proceedings of the 15th annual conference on Genetic and evolutionary computation, 2013: 1229-1236.

[39] Shahroudy Amir, Wang Gang, Ng Tian-Tsong. Multimodal feature fusion for action recognition in rgb-d sequences [C]//Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on, 2014: 1-4.

[40] Zhu Yu, Chen Wenbin, Guo Guodong. Fusing multiple features for depth-based action recognition[J]. ACM Trans. Intell. Syst. Technol., 2015, 6(2): 1-20.

[41] Kobayashi Takumi, Otsu Nobuyuki. Motion recognition using local auto-correlation of space-time gradients [J]. Pattern Recognition Letters, 2012, 33(9): 1188-1195.

[42] Cho Kyunghyun, Chen Xi. Classifying and visualizing motion capture sequences using deep neural networks [J]. arXiv preprint arXiv:1306.3874, 2013.

[43] Wang Pichao, Li Wanqing, Gao Zhimin, et al. Deep convolutional neural networks for action recognition using depth map sequences[J]. arXiv preprint arXiv: 1501.04686, 2015.

[44] Müller Meinard, Baak Andreas, Seidel Hans-Peter. Efficient and robust annotation of motion capture data [C]//Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2009: 17-26.

[45] Müller Meinard, Röder Tido, Clausen Michael, et al. Documentation mocap database hdm05[R]. Technical report, No. CG-2007-2, Universität Bonn, June 2007.

[46] Fothergill Simon, Mentis Helena, Kohli Pushmeet, et al. Instructing people for training gestural interactive systems[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012: 1737-1746.

[47] Seidenari Lorenzo, Varano Vincenzo, Berretti Stefano, et al. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, 2013: 479-485.

[48] Kurakin Alexey, Zhang Zhengyou, Liu Zicheng. A real time system for dynamic hand gesture recognition with a depth sensor [C]//Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, 2012: 1975-1979.

[49] Ni Bingbing, Wang Gang, Moulin Pierre. Rgbd-hudaact: A color-depth video database for human daily activity recognition[C]//2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011: 1147-1153.

[50] Jaeyong Sung, Ponce C, Selman B, et al. Unstructured human activity detection from rgbd images[C]//Robotics and Automation (ICRA), 2012 IEEE International Conference on, 2012: 842-849.

[51] Ellis Chris, Masood Syed Zain, Tappen Marshall F, et al. Exploring the trade-off between accuracy and observational latency in action recognition[J]. International Journal of Computer Vision, 2013, 101(3): 420-436.

(责任编辑 王卫勋)