人体行为识别数据集研究进展

朱红蕾 [1] 朱昶胜 [1] 徐志刚 [1]

摘 要 人体行为识别是计算机视觉领域的一个研究热点, 具有重要理论价值和现实意义. 近年来, 为了评价人体行为识别方法的性能, 大量的公开数据集被创建. 本文系统综述了人体行为识别公开数据集的发展与前瞻: 首先, 对公开数据集的层次与内容进行归纳. 根据数据集的数据特点和获取方式的不同, 将人体行为识别的公开数据集分成 4 类. 其次, 对 4 类数据集分别描述, 并对相应数据集的最新识别率及其研究方法进行对比与分析. 然后, 通过比较各数据集的信息和特征, 引导研究者选取合适的基准数据集来验证其算法的性能, 促进人体行为识别技术的发展. 最后, 给出公开数据集未来发展的趋势与人体行为识别技术的展望.

关键词 计算机视觉, 行为识别, 真实场景, 多视角, 多模态

引用格式 朱红蕾, 朱昶胜, 徐志刚. 人体行为识别数据集研究进展. 自动化学报, 2018, **44**(6): 978−1004

DOI 10.16383/j.aas.2018.c170043

Research Advances on Human Activity Recognition Datasets

ZHU Hong-Lei[1] ZHU Chang-Sheng[1] XU Zhi-Gang[1]

Abstract Human activity recognition is an important research field of computer vision with important theoretical value and practical significance. In recent years, a large number of public datasets have been created for evaluation of human activity recognition methodologies. This paper reviews the progress and forecast the future of public datasets for human activity recognition. First, the hierarchy and contents of the public datasets are summarized, and the public datasets are divided into four categories according to the characteristics and acquiring methods. Then, the four categories are described and analyzed separately. Meantime, the state-of-the-art research results and corresponding methods of the public datasets are introduced to researchers. By comparing the information and characteristics of each dataset, researchers can be guided in the selection of the most suitable dataset for benchmarking their algorithms, so as to promote the technology progress of human activity recognition. Finally, the future trends of the public datasets and the prospects of human activity recognition are discussed.

Key words Computer vision, activity recognition, real scenes, multi-view, multimodality

Citation Zhu Hong-Lei, Zhu Chang-Sheng, Xu Zhi-Gang. Research advances on human activity recognition datasets. *Acta Automatica Sinica*, 2018, **44**(6): 978−1004

人体行为识别是一个多学科交叉的研究方向, 涉及图像处理、计算机视觉、模式识别、机器学习、人工智能等多个学科, 是计算机视觉领域的一个重要研究课题[1]. 随着数字图像处理技术和智能硬件制造技术的飞速发展, 人体行为识别在智能视频监控[1−3]、自然人机交互[4−6]、智能家居[7−9]、虚拟现实[10] 等领域具有广泛的应用前景.

自以色列魏茨曼科学研究所于 2001 年发布基于事件的视频分析数据库[11] 以来, 许多人体行为数据集陆续公开发布, 对促进人体行为识别方法的研究起到关键的作用, 也对计算机视觉研究的发展具有很大的推动作用. 公开的人体行为数据集为众多研究者提供了一定的研究规范, 使研究者可以利用相同的输入数据来比较不同识别方法的相关性能, 是校验识别方法性能优劣的重要标准.

人体行为数据集的更新和发展在计算机视觉领域起到了方向标的作用. 而各个公开的人体行为数据集在相机状态、拍摄视角、活动场景、行为类别以及视频规模等方面具有很大的差异. 因此, 对公开数据集进行对比分析, 有利于研究者根据自己的需求选择合适的数据集, 缩短研究周期. 截至目前, 已有一些涉及行为识别数据集相关的综述性文章[12−14]. Ahad 等[12] 简单介绍了与人体行为相关的数据集信息. Chaquet 等[13] 较详尽地介绍与人体行为和活动相关的数据集, 并罗列出应用各个数据集的相应文献, 但没有提供数据集的最新研究成果. 而 Zhang

收稿日期 2017-01-16 录用日期 2017-07-18
Manuscript received January 16, 2017; accepted July 18, 2017
国家自然科学基金 (61563030), 甘肃省自然科学基金 (1610RJZA027) 资助
Supported by National Natural Science Foundation of China (61563030), and Natural Science Foundation of Gansu Province (1610RJZA027)
本文责任编委 桑农
Recommended by Associate Editor SANG Nong
1. 兰州理工大学计算机与通信学院 兰州 730050
1. School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050

等[14] 针对涉及深度信息的行为识别数据集进行了详细介绍, 但有些数据集的信息和研究成果需要更新. 还有一些综述性文章[15−18], 侧重于行为识别的研究方法, 而对相关数据集介绍较简单.

根据数据集的数据特点和获取方式, 可以把人体行为识别领域常用的公开数据集分为 4 类: 通用数据集、真实场景数据集、多视角数据集和特殊数据集. 而根据人的行为方式可以将公开数据集分为三类: 个体行为数据集、交互行为数据集和群体行为数据集. 根据第一种分类方式, 下面的章节将分别对 4 类数据集及其研究方法进行详细介绍.

1　通用数据集

通用行为识别数据集, 它们包含受试者在受限场景下执行的一系列简单动作, 如 KTH[19] 和 Weizmann[20−21].

KTH[19] 数据集发布于 2004 年, 是计算机视觉领域的一个里程碑. 该数据集提供了 4 类场景下 25 个不同受试者的 6 种人体行为: 步行 (Walking)、慢跑 (Jogging)、跑步 (Running)、拳击 (Boxing)、挥手 (Hand waving) 和拍手 (Hand clapping), 其示例如图 1 所示.

该数据集的 4 类场景分别为室外 (s1)、室外不同着装 (s2)、室外放大 (s3) 和室内 (s4). 数据集一共包含 599 个视频, 其中 8 个受试者的视频作为训练集, 8 个受试者的视频作为验证集, 9 个受试者的视频作为测试集. 该数据集的视频具有尺度、衣着和光照的变化, 但其场景中背景相对静止, 摄像机位置也相对固定, 只有焦距的变化. 因此该数据集相对比较简单, 但由于场景变化, 目前其识别准确率未能达到 100 %. Zhou 等[22] 基于多核学习 (Multiple kernel leaning, MKL), 针对时空兴趣点 (Space-time interest points, STIP) 利用语义上下文特征树模型增强行为描述符的辨别力, 其识别率达到 98.67 %. 而 Xu 等[23] 利用三个低层特征: STIP、空间星图 (SSG) 和时间星图 (TSG) 构建基于中层特征的视觉词袋 (MLDF), 达到 98.83 % 的识别率.

Weizmann[20−21] 数据集发布于 2005 年, 一共包含 9 个不同受试者的 10 种人体行为: 走 (Walk)、跑 (Run)、双腿跳 (Jump)、侧身跑 (Gallop sideways)、弯腰 (Bend)、挥单手 (One-hand wave)、挥双手 (Two-hands wave)、原地跳 (Jump in place)、开合跳 (Jumping Jack) 和单腿跳 (Skip). 该数据集一共包含 93 个视频, 其分辨率较低, 为 144×180. 数据集视频场景中的背景、视角及摄像头都是静止的, 并提供利用背景消减法得到的剪影信息, 如图 2 所示. 此外, 该数据集还提供包含两个单独动作的

图 1　KTH 数据集示例图[19]

Fig. 1　Sample images of KTH dataset[19]

图 2　Weizmann 数据集示例及其剪影图[24]

Fig. 2　Sample images and silhouettes of Weizmann dataset[24]

视频序列: 一个是不同视角下人体行走的视频; 另一个为衣着和人物等方面有细微差异的行走动作序列. 该数据集比较简单, 研究者于 2008 年利用度量学习方法[25] 和中层运动特征[26] 已达 100 % 的识别率.

通用数据集提出较早, 包含行为类型简单、规模较小. 目前研究者对其关注较少, 仅利用它来对比验证算法的性能.

通用数据集中各数据集的最新识别率、研究方法、评价方案等信息如表 1 所示.

2 真实场景数据集

真实场景数据集主要是从电影或视频中收集的数据, 比如 Hollywood[27]、UCF Sports[28]、Hollywood 2[29]、UCF YouTube[30]、Olympic Sports[31]、HMDB51[32]、UCF50[33]、UCF101[34]、Sports-1M[35] 数据集等. 它们共同的特点是相机、场景不固定且同类动作的类内散度比较大, 因而极具挑战性.

Hollywood (HOHA)[27] 数据集来自 32 部电影, 从中抽取由不同的演员在不同的环境下执行的相同动作. 该数据集包括 8 种行为类别: 接电话 (AnswerPhone)、下车 (GetOutCar)、握手 (HandShake)、拥抱 (HugPerson)、亲吻 (Kiss)、坐下 (SitDown)、端坐 (SitUp)、起立 (StandUp), 并具有一个或多个标签. 该数据集被划分成两部分: 从 12 部电影获得的 2 个训练集和从其余的 20 部电影获得的测试集. 其中, 2 个训练集包括一个自动训练集和一个干净训练集. 自动训练集使用自动脚本进行行为标注, 包含 233 个视频样本, 并具有超过 60 % 的正确标签; 而干净训练集则包含 219 个视频样本, 具有手动验证标签. 测试集包含 211 个视频样本, 均具有手动验证标签. Kulkarni 等[36] 针对连续行为识别, 基于动态时间规整提出一种新颖的视觉对准技术动态帧规整 (DFW), 达到 59.9 % 的识别率. 而 Shabani 等[37] 基于标准判别词袋行为识别框架, 通过对比基于结构的特征和基于运动的特征的性能, 使用非对称运动特征进行有效的稀疏紧凑表示达到 62 % 的识别率.

Hollywood 2[29] 数据集是 Hollywood[27] 的扩展, 来自 69 部电影, 包含 12 种行为类别和 10 类场景, 共有 3 669 个视频. 该数据集包含两个子集: 行为数据集 (2 517 个视频, 现实际有 2 442 个视频) 和场景数据集 (1 152 个视频). 行为数据集 (Actions) 在 Hollywood[27] 的基础上增加了 4 种行为类别: 开车 (DriveCar)、吃饭 (Eat)、打架 (FightPerson) 和跑 (Run), 其示例如图 3 所示.

该数据集的训练集从 33 部电影中获得, 而测试集从其余的 36 部电影中获得. 行为数据集包含 2 个训练集和一个测试集 (884 个视频). 其中, 2 个训练集包括一个自动训练集和一个干净训练集. 自动训练集使用自动脚本进行行为标注, 包含 810 个视频样本 (现实际有 735 个); 而干净训练集则包含 823 个视频样本. 场景数据集 (Scenes) 包含一个自动标注的训练集 (570 个视频) 和一个测试集 (582 个视频). 因为视频中演员的表情、姿态、穿着各异, 再加上相机运动、光照条件、遮挡、背景等诸多因素影响, 其视频接近于真实场景下的情况, 因此该数据集极具挑战性. Fernando 等[38] 采用卷积神经网络 (Cellular neural networks, CNN), 利用 Fisher 向量 (Fisher vector, FV) 和秩池化 (Rank pooling, RP) 对改进稠密轨迹 (Improved dense trajectory, iDT)[39] 描述符编码, 并结合分层秩池化 (HRP) 编码的 CNN 特征, 达到 76.7 % 的识别率. Liu 等[40] 提出一种分层聚类多任务学习 (HC-MTL) 方法, 同时利用低秩 (Low rank) 和组稀疏 (Group sparsity) 结构进行正则化, 达到 78.5 % 的识别率. 而 Wang 等[41] 利用改进的双流卷积神经网络 (Two-stream ConvNets, TCNN)[42], 在多

表 1 通用数据集的最新研究成果概览表

Table 1 Summary of state-of-the-art research results on general datasets

数据集名称	最新识别率	年份	研究方法	评价方案
KTH	98.83 %[23]	2016	MLDF	CS: Tr: 16; Te: 9
	98.67 %[22]	2016	Semantic context feature-tree (MKL)	CS: Tr: 16; Te: 9
	98.5 %[43]	2015	Local region tracking (HBRT/VOC)	CS: Tr: 16; Te: 9
Weizmann	100 %[44]	2017	3D-TCCHOGAC+3D-HOOFGAC	LOOCV
	100 %[45]	2016	$\Re$ transform + LLE (SVM)	LOOCV
	100 %[46]	2016	SDEG + $\Re$ transform	LOOCV
	100 %[47]	2014	3D cuboids + mid-level feature (RF)	LOSOCV
	100 %[25]	2008	Metric learning	LOSOCV
	100 %[26]	2008	Mid-level motion features	LOOCV

*Tr: training set; Te: test set; CS: cross-subject; LOOCV: leave-one-out cross validation; LOSOCV: leave-one-subject-out cross validation

图 3　Hollywood 2 数据集示例图[48]

Fig. 3　Sample images of Hollywood 2 Dataset[48]

个卷积层计算 EPT (Evolution-preserving dense trajectory) 描述符, 并与稠密轨迹 (DT)[49] 描述符融合, 同时利用 VideoDarwin 技术, 达到 78.6 % 的识别率.

UCF Sports[28] 数据集主要来自 BBC 和 ESPN 等广播电视频道, 包含 150 个视频. 该数据集包含 10 种运动类别: 跳水 (Diving)、高尔夫挥杆 (Golf Swing)、踢足球 (Kicking)、举重 (Lifting)、骑马 (Riding Horse)、跑步 (Running)、滑板 (Skateboarding)、平衡木 (Swing-Bench)、双杠 (Swing-Side) 和行走 (Walking), 其示例如图 4 所示.

图 4　UCF Sports 数据集示例图[50]

Fig. 4　Sample images of UCF Sports Dataset[50]

该数据集的视频具有较高分辨率, 是各种现实场景的自然行为, 因此其在动作类型、相机运动、视角、光照和背景等方面有较大差异, 具有一定的挑战性, 并有助于研究不受约束环境的行为识别. 目前, Tong 等[44] 提出 3D-TCCHOGAC 和 3D-HOOFGAC 两个构建动态描述符的方法, 并利用这两个动态描述符与静态描述符融合形成一种行为识别新框架, 达到 96 % 的识别率. 而 Harbi 等[43] 有别于传统的基于时空兴趣点技术, 通过先进的人体检测和分割方法 (HBRT/VOC) 提取时空人体区域信息, 利用局部约束线性编码 (LLC) 达到 96.2 % 的识别率.

UCF YouTube[30] 数据集目前被称为 UCF11, 是由中佛罗里达大学 (University of Central Florida, UCF) 计算机视觉研究中心发布的, 包含 1 600 个视频. 该数据集共有 11 种行为类别: 篮球投篮 (b_shooting)、骑自行车 (cycling)、跳水 (diving)、高尔夫挥杆 (g_swinging)、骑马 (r_riding)、足球颠球 (s_juggling)、荡秋千 (swinging)、打网球 (t_swinging)、跳蹦床 (t_jumping)、排球扣球 (v_spiking)、与狗一起散步 (g_walking), 其示例如图 5 所示.

该数据集的视频格式是 MPEG 格式, 对于每个类别的视频被分成 25 组, 每组至少 4 个行为视频. 同一组的视频具有一些共同的特征, 如演员相同、背景相似、视角相似等. 因此, 虽然该数据集也具有相机运动、视角、背景复杂度、光照条件等变化, 但由于类内相似度较高, 目前其识别准确率较高. Peng 等[51] 通过在表征层将传统的 Fisher 向量与堆叠 Fisher 向量 (SFV) 合并, 达到 93.77 % 的识别率. Liu 等[52] 提出一个深度学习框架 CNRF, 采用时空 CNN 从原始输入帧学习不变特征, 同时采用结合条件随机场 (CRF) 的 CNN 捕获输出之间的相互依赖关系, 通过联合学习它们的参数, 达到 94.4 % 的识别率. Sun 等[53] 利用词袋量化将残差向量压缩成低维残差直方图, 并与多个迭代高阶残差向量生成的高阶残差直方图连接形成分层词袋模型 (HBoW), 然后采用内部归一化处理, 达到 94.50 % 的识别率.

Olympic Sports[31] 数据集来自于 YouTube, 包含运动员练习的 783 个视频. 该数据集包含 16 种运动类别: 跳高 (high-jump)、跳远 (long-jump)、三级跳远 (triple-jump)、撑杆跳 (pole-vault)、单手上篮

(basketball lay-up)、打保龄球 (bowling)、网球发球 (tennis-serve)、10 米跳台 (platform)、铁饼 (discus)、链球 (hammer)、标枪 (javelin)、铅球 (shot put)、3 米跳板 (springboard)、举重抓举 (snatch)、举重挺举 (clean-jerk) 和跳马 (vault), 其示例如图 6 所示. 该数据集在亚马逊土耳其机器人的帮助下注释其类标签, 包含复杂运动、严重遮挡、相机运动等因素影响. 目前, Sekma 等[54] 基于人体检测的 iDT 描述符提出一种多层 Fisher 向量编码的方法, 达到 96.5 % 的识别率. 而 Li 等[55] 通过深度卷积神经网络 (DCNN) 获得短时动态特征; 利用线性动态系统 (LDS) 得到中间范围动态特征; 借助局部特征聚合描述符 (VLAD) 获得长期的不均匀动态特征, 并在考虑上述不同级别视频动态特征的基础上提出 VLAD^3 表征方法, 同时结合 iDT 描述符进一步提高性能, 获得 96.6 % 的识别率.

图 5 UCF YouTube 数据集示例图[30]

Fig. 5 Sample images of UCF YouTube Dataset[30]

图 6 Olympic Sports 数据集示例图

Fig. 6 Sample images of Olympic Sports Dataset

HMDB51[32] 数据集主要来源于电影, 只有一小部分来自公共数据库, 如 Prelinger 存档、YouTube 和 Google 视频. 该数据集包含 6 849 个视频, 分为 51 种行为类别, 每种行为包含至少 101 个视频. 该数据集的行为类别可以归纳为 5 种类型: 1) 普通面部动作: 微笑、大笑、咀嚼、说话; 2) 操纵对象的面部动作: 抽烟、吃、喝; 3) 普通身体运动: 侧手翻、拍手、攀登、爬楼梯、俯冲、落地、反手空翻、倒立、跳、引体向上、俯卧撑、跑、坐下、仰卧起坐、翻筋斗、站起来、转身、走、挥手; 4) 与对象交互的身体运动: 梳头、抓球、拔剑、运球、打高尔夫、打东西、踢足球、捡东西、倒东西、推东西、骑自行车、骑马、投篮、射箭、射枪、打球棒、练剑、扔东西; 5) 与人交互的身体运动: 击剑、拥抱、踢人、亲吻、拳击、握手、斗剑, 其部分示例如图 7 所示. 因为该数据集来源不同, 并伴有遮挡、相机移动、复杂背景、光照条件变化等诸多因素影响, 导致其识别准确率较低, 极具有挑战性. 最初, 该数据集的识别率为 23.18 %[32]. 2016 年, Feichtenhofer 等[56] 利用 TCNN 融合时间和空间特征达到 69.2 % 的识别率. Wang 等[57] 则利用三个 TCNN 构建时间分割网络 (TSN) 达到 69.4 % 的识别率, 略高于前者. 此外, Wang 等[58] 研究了行为和场景之间的关系, 通过深度卷积神经网络 Places205-VGGNet[59] 模型获得场景特征, 同时利用静态场景编码和动态场景编码作为场景特征的补充, 再与运动特征结合, 将识别率提高到 73.6 %.

UCF50[33] 数据集来自 YouTube 的现实视频, 是 UCF11[30] 的扩展. 该数据集的行为类别由 11 种扩展到 50 种, 包含 6 676 个视频 (现在实际有 6 681 个). 该数据集增加的 39 种行为类别为: 棒球投掷、卧推、台球击球、蛙泳、挺举、击鼓、击剑、弹吉他、跳高、赛马、呼啦圈、掷标枪、杂耍球、跳绳、开合跳、皮划艇、弓步、阅兵、调糊、双截棍、弹钢琴、扔披萨、撑竿跳、鞍马、引体向上、拳击、俯卧撑、室内攀岩、爬绳、赛艇、萨尔萨舞旋转、滑板、滑雪、摩托艇、打手鼓、太极、掷铁饼、弹小提琴和溜溜球, 其示例如图 8 所示. 每种行为类别也包含 25 组, 每组包含 4～23 个视频, 具有一些共同的特征. 因此, 该数据集识别率较高. 截至目前, Lan 等[60] 为解决高斯金字塔不能在粗尺度产生新特征的问题, 提出一种新的特征增强技术 MIFS. MIFS 使用一系列差分滤波器提取堆叠特征, 通过多次时间跳跃参数化, 实现频率空间的平移不变性, 同时以粗尺度重新获取的信息来补偿使用差分算子丢失的信息, 提高基于差分滤波器特征的可学习性, 达到 94.4 % 的识别率. Ijjina 等[61] 利用遗传算法和深度卷积神经网络, 采用 5 折交叉验证达到 99.98 % 的识别率.

UCF101[34] 数据集又是 UCF50[33] 的扩展, 包含 101 种动作类别, 共计 13 320 个视频片段. 该数据集的行为类别可以分成 5 类: 1) 人与对象的交互; 2) 身体运动; 3) 人之间的交互; 4) 乐器演奏; 5) 体育运动, 其部分示例如图 9 所示. 该数据集的每种行为类别包含 25 组, 每组包含 4～7 个视频片段. 该数据集由用户上传, 来自于无约束的现实环境, 平均剪辑长度为 7.21 秒, 包含相机运动、杂乱背景、不同光照条件、遮挡、低质量等不确定因素, 因此该数据集非常具有挑战性, 也引起了众多研究者的关注. 2012 年最初的识别率为 43.9 %[34]. 2016 年, Feichtenhofer 等[56] 利用 TCNN 将识别率提升到 93.5 %. 同年, Lev 等[62] 基于 FV, 利用递归神经网络 (RNN) 生成概率模型, 同时利用反向传播算法 (BP) 计算偏导数, 达到 94.08 % 的识别率. 而 Wang 等[57] 则利用 TSN 进一步将识别率提升到 94.2 %, 给研究者提供了更好的研究思路.

(a) 普通面部动作
(a) General facial actions

(b) 操纵对象的面部动作
(b) Facial actions with object manipulation

(c) 普通身体运动
(c) General body movements

(d) 与对象交互的身体运动
(d) Body movements with object interaction

(e) 与人交互的身体运动
(e) Body movements for human interaction

图 7　HDMB51 数据集示例图

Fig. 7　Sample images of HDMB51 dataset

THUMOS 挑战开始于 2013 年, 基于 UCF101[34] 数据集, 其目的是对含有大量类别的真实原始视频的大规模行为识别探索新的方法. THUMOS′13[63] 的基准数据集在 UCF101 数据集的基础上增加了 24 类的注释框, 其中 14 个类来自 UCF101, 10 个类来自 UCF11. THUMOS′14[64] 的基准数据集在 THUMOS′13 的基础上增加了 2 500

图 8 UCF50 数据集示例图[33]

Fig. 8 Sample images of UCF50 dataset[33]

个背景视频、1 010 个验证视频和 1 574 个测试视频. THUMOS′15[65] 的基准数据集是 THUMOS′14 数据集的扩展, 增加到 2 980 个背景视频、2 104 个验证视频和 5 613 个测试视频. 而且 THUMOS 增加的视频是未经修剪的原始视频, 其中还包括验证集和测试集中每种行为的负背景视频, 使行为识别任务更加困难. 在 2015 年的 THUMOS 挑战赛中, 参赛组大都采用深度学习技术, 利用 VGG-Net 或 CNN 模型进行研究和改进, 其中悉尼科技大学和美国卡内基梅隆大学的联合参赛组取得 74.6 %[66] 的

最好识别准确率. 而后 Li 等[55] 提出融合不同级别视频动态特征的 VLAD^3 表征方法, 同时利用 iDT 描述符获得 80.8 % 的识别率.

图 9　UCF101 数据集示例图

Fig. 9　Sample images of UCF101 dataset

Sports-1M[35] 数据集是 Google 公布的一个大型视频数据集, 来自于公开的 YouTube 视频. 该数据集包含 487 种体育运动项目, 共计 1 133 158 个视频. 该数据集中每种行为类别包含 1 000 ∼ 3 000 个视频, 其中有大约 5 % 的视频带有多个标注. 该数据集包含的体育运动项目可以分为 6 大类: 水上运动、团队运动、冬季运动、球类运动、对抗运动、与动物运动. 而且各类别在叶级层次差异很小, 如包含 6 个不同类型的保龄球和 23 个不同类型台球等. 自数据集创建以来, 约有 7 % 的视频已经被用户删除. 由于该数据集来自公开视频, 所以相机运动不受限制, 导致光流参数在视频间变化较大, 给视频的识别带来一定的困难. 目前, Mahasseni 等[67] 基于深度卷积神经网络 (DCNN) 和两层长短时记忆 (LSTM) 的多层体系结构, 同时利用 3D 骨架序列补充训练数据特征来改进大规模行为识别的效率. 在正则化约束参数 $g1$ 下, Hit@1 的识别率为 73.4 %, Hit@5 的识别率为 91.3 %; 而在正则化约束参数 $g3$ 下, Hit@1 的识别率为 75.9 %, Hit@5 的识别率为 91.7 %.

真实场景数据集的行为类别、数据规模、场景复杂度不断增大, 给研究者提出了新挑战. 而随着近年来深度学习在机器视觉领域的研究与应用, 研究者基于深度学习技术, 利用不同的模型, 如卷积神经网络 (CNN)、深度卷积神经网络 (DCNN)、递归神经网络 (RNN)、双流卷积神经网络 (TCNN) 等, 同时结合不同的方法使相关数据集的识别率有了较大地提升.

真实场景数据集中各数据集的最新识别率、研究方法、评价方案等信息如表 2 所示.

3　多视角数据集

视频行为分析最大的困难之一是由视角变化引起的特征不确定性. 多视角数据集为视角变化情况下研究行为的旋转不变性提供了基准数据集. 常见的多视角数据集有: IXMAS[68]、MuHAVi[69]、PETS[70−71] 等.

INRIA Xmas Motion Acquisition Sequence (IXMAS)[68] 数据集是由法国国家信息与自动化研究所 (Institute for Research in Computer Science and Automation, INRIA) 发布的, 是多视角和三维研究的重要校验基石. 该数据集是从 5 个视角拍摄的, 室内的 4 个方向和顶部的 1 个方向. 目前, 该数据集更新至总共由 12 个受试者完成 13 种不同的日常行为, 共计 180 个视频. 该数据集的 13 种日常行为: 看手表、抱胳膊、抓头、坐下、起来、转身、走、挥手、拳击、踢、指、捡和扔, 其同一动作 5 个视角的示例及其剪影如图 10 所示. 其中扔的动作又可以细分为两类: 过头扔和从下方扔.

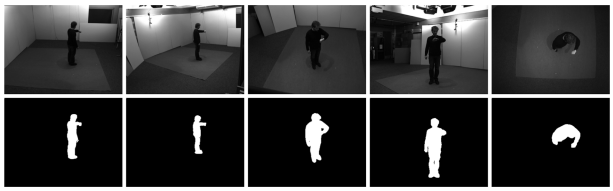


图 10　IXMAS 数据集同一动作的 5 个视角及其剪影示例图

Fig. 10　Sample images and the corresponding silhouettes for the same action of IXMAS dataset (5 cameras)

该数据集的视频中受试者顺序执行 13 种日常行为动作, 并重复执行 3 次. 而最早公开的数据集[68] 仅包含 10 个受试者执行的 11 种日常行为, 比目前公开的数据集少了 2 个受试者和两种行为 (指和扔). 另外, 该数据集还提供人体轮廓和体积元等信息. 该数据集非常具有挑战性, 虽然摄像机是固定的, 环境的光照条件和背景基本不变, 但是受试者可以自由选择自己的位置和姿态, 故存在较大的外观变化、内部类变化和遮挡问题. 针对该数据集的特点, 研究者分别从单视角和多视角两个方面进行研究. 对常见单视角的 5 种行为 (看手表、抱胳膊、抓头、坐下和起来), Ashraf 等[72] 利用对极几何单

应性的一致性, 将身体姿态看作 11 个身体点研究视角无关的行为识别, 其识别率为 91.6 %. 而对单视角的 11 种行为, Ji 等[73] 通过连接相邻视点空间之间的子行为模型建立多视角转换隐马尔科夫模型 (HMM), 达到 92.7 % 的识别率. 对 5 个视角的 11 种行为, Gao 等[74] 通过有监督迁移字典对学习, 利用 Cuboid 特征获得 95.3 % 的识别率; 利用 STIP 特征获得 95.1 % 的识别率. 而 Wu 等[75] 利用基于多视角最大间距的支持向量机 (MMM-SVM), 达到 95.54 % 的识别率.

多视角 MuHAVi[69] 数据集最早由英国工程和物理科学研究委员会 (EPSRC) 项目支持, 而目前则由智力科学技术研究委员会 (CONICYT) 常规项目支持. 该数据集由 7 个受试者执行, 包含 8 个视角 (其位置如图 11 所示), 共计 952 个视频. 图 11 的中间区域是行为执行区域, 在现场地板用白色胶带标记.

该数据集包含 17 种行为类别: 来回走、跑步停

表 2 真实场景数据集的最新研究成果概览表

Table 2 Summary of state-of-the-art research results on real scene datasets

数据集名称	最新识别率	年份	研究方法	评价方案
Hollywood	62 %[37]	2012	Asymmetric motions (BoW)	Tr: 219 vedios; Te: 211vedios
	59.9 %[36]	2015	DFW (BoW)	Tr: 219 vedios; Te: 211vedios
	56.51 %[76]	2016	STG-MIL	Tr: 219 vedios; Te: 211vedios
Hollywood 2	78.6 %[41]	2017	EPT + DT + VideoDarwin (TCNN)	Tr: 823 videos; Te: 884 videos
	78.5 %[40]	2017	HC-MTL + L/S Reg	Tr: 823 videos; Te: 884 videos
	76.7 %[38]	2016	HRP + iDT (VGG-16)	Tr: 823 videos; Te: 884 videos
UCF Sports	96.2 %[43]	2015	Local region tracking (HBRT/VOC)	all classes
	96 %[44]	2017	3D-TCCHOGAC + 3D-HOOFGAC	LOOCV
	95.50 %[47]	2014	3D cuboids + mid-level feature (RF)	LOOCV
UCF YouTube	94.50 %[53]	2016	HboW	LOOCV
	94.4 %[52]	2016	CNRF (CNN)	LOVOCV
	93.77 %[51]	2014	FV + SFV	LOGOCV
Olympic Sports	96.60 %[55]	2016	VLAD^3 + iDT (CNN)	each class video: Tr: 40; Te: 10
	96.5 %[54]	2015	iDT + HD (multi-layer FV)	not mentioned
	93.6 %[77]	2017	Bag-of-Sequencelets	Tr: 649 videos; Te: 134 videos
HMDB51	73.6 %[58]	2016	scene + motion (DCNN)	three train/test splits
	69.40 %[57]	2016	TSN (TCNN)	three train/test splits
	69.2 %[56]	2016	spatiotemporal fusion (TCNN)	three train/test splits
UCF50	99.98 %[61]	2016	GA (CNN)	5-fold cross-validatin
	94.4 %[60]	2015	MIFS	LOGOCV
	94.1 %[78]	2013	weighted SVM	5-fold LOGOCV
UCF101	94.20 %[57]	2016	TSN (TCNN)	three train/test splits
	94.08 %[62]	2016	RNN-FV (C3D + VGG-CCA) + iDT	three train/test splits
	93.5 %[56]	2016	spatiotemporal fusion (TCNN)	three train/test splits
THUMOS′15	80.8 %[55]	2016	VLAD^3 + iDT (CNN)	5-fold cross-validation
	76.8 %[55]	2016	VLAD^3 (CNN)	5-fold cross-validation
	74.6 %[66]	2015	VLAD + LCD (VGG-16)	5-fold cross-validation
	70.0 %[79]	2015	Stream Fusion + Linear SVM (VGG-19)	Tr: UCF101 dataset; Te: val15
	65.5 %[80]	2015	iDT + LCD + VLAD (VGG-16)	Tr: UCF101 dataset; Vs: val15 Te: UCF101 dataset + val15
Sports-1M (Hit@1)	75.9 %[67]	2016	RLSTM-$g3$ (GoogLeNet)	not mentioned
	73.4 %[67]	2016	RLSTM-$g1$ (GoogLeNet)	not mentioned
	73.10 %[81]	2015	LSTM on Raw Frames LSTM on Optical Flow (GoogLeNet)	1.1 million videos

*LOVOCV: leave-one-video-out cross validation; LOGOCV: leave-one-group-out cross validation; Vs: validation set

止、拳击、踢、强迫倒、拉重物、捡起扔物体、步行摔倒、看车、膝盖爬行、挥胳膊、画涂鸦、跳过栅栏、醉走、爬梯子、打碎物体、跳过间隙, 其 8 个视角的示例如图 12 所示. 针对 4 个视角, Moghaddam 等[82] 利用基于轮廓的扇形极值点, 采用 HMM 进行分类, 达到 92.1 % 的识别率; 而 Wu 等[83] 提出视角无关的 LKSSVM 学习算法, 达到 97.48 % 的识别率. Alcantara 等[84] 针对所有视角, 利用累积运动形状 (CMS) 和多层描述符, 采用多级 K 近邻法 (K-NN) 进行分类, 达到 91.6 % 的识别率.

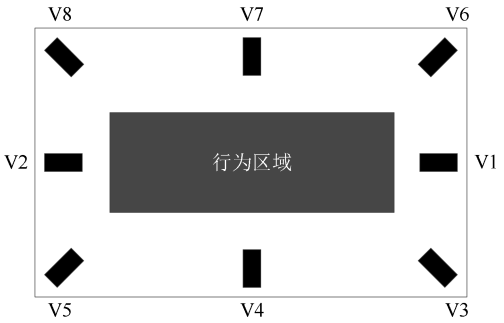


图 11　8 个摄像机配置的顶视图[69]

Fig. 11　The top view of the configuration of 8 cameras[69]

多视角 MuHAVi-MAS[69] 数据集是 MuHAVi[69] 的子集, 并对轮廓数据进行了手动标注. 该数据集由 2 个受试者执行, 仅包含侧面和 45° 两个视角 (位置如图 11 中所示的 V3 和 V4), 共计 136 个视频. 该数据集的行为划分更加精细, 一共含有 14 种行为 (MuHAVi-14): 向左倒、向右倒、自卫踢、自卫拳击、右踢、右击、从左向右跑、从右向左跑、从左边站起来、从右边站起来、从左向后转、从右向后转、从左向右走和从右向左走, 其两个视角的部分行为剪影如图 13 所示. 由于该数据集中包含视角变化, 行为类别之间具有较大的混淆性, 如从左向右跑和从右向左跑都可以视为跑, 因此, 具有一定的挑战性. Chaaraoui 等[85] 利用低维径向概括特征 (Radial summary feature) 和特征子集选择 (Feature subset selection) 进行特征级优化, 达到 98.5 % 的识别率. 而 Cai 等[86] 利用姿势字典学习达到 98.53 % 的识别率.

另外, MuHAVi-14 的 14 种原始行为也可以合并为 8 种 (MuHAVi-8): 倒 (向左/右)、站起来 (从左/右)、右踢、右击、自卫 (踢/拳击)、跑 (向左/右)、走 (向左/右) 和向后转 (从左/右). 该数据集由于合并混淆性行为而降低了识别难度, Chaaraoui 等[85]、Chaaraoui 等[87] 和 Alcantara 等[84, 88] 都实现了 100 % 的识别率.

PETS (International Workshop on Performance Evaluation of Tracking and Surveillance), 其全称为跟踪与监控性能评估会议. 该会议自 2000 年在法国召开第一届以来, 截至 2016 年, 已举行了 16 届. 它的数据集是从现实生活中获取的, 主要来源于直接从视频监控系统拍摄的视频. PETS 研讨会的目标是通过提供基准数据集来促进计算机视觉中检测和跟踪技术的发展.

PETS 2009[70] 的基准数据集采自在英国雷丁大学的 Whiteknights 校区, 涉及大约 40 个受试者, 有 8 个摄像机位于不同角度进行拍摄, 其位置和方向的平面图如图 14 所示, 而实景拍摄示例如图 15 所示. 该数据集记录了不同的人群活动序列, 分为三个数据集: 数据集 S1 涉及人群人数和密度估计; 数据集 S2 用于人群中个体的跟踪; 数据集 S3 涉及人群流分析和事件检测.

PETS 2014[71] 的基准数据集由欧盟项目 ARENA 赞助, 称为 “ARENA 数据集”. 该数据

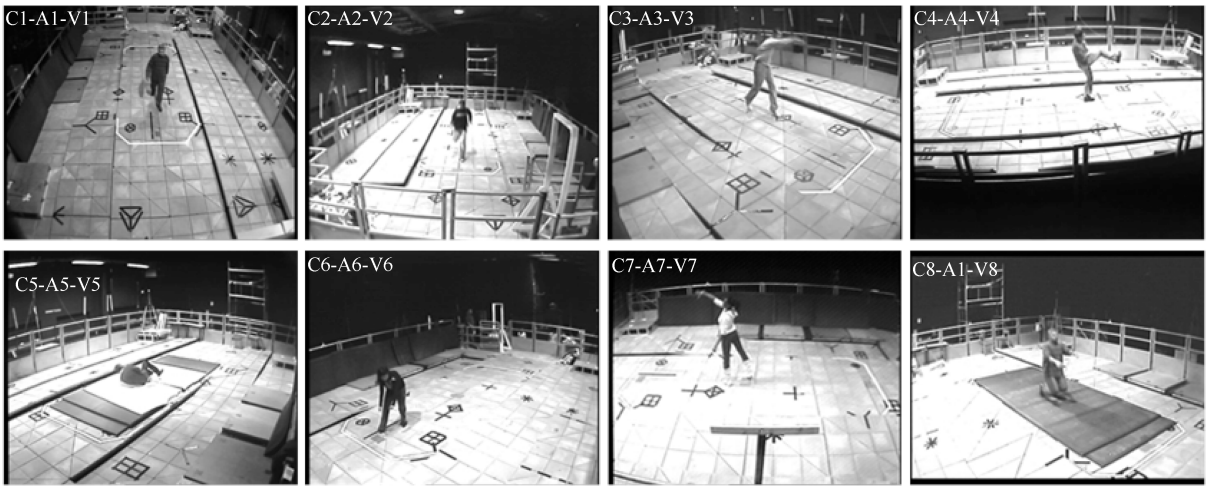


图 12　MuHAVi 数据集的 8 个视角示例图[69]

Fig. 12　Sample images of MuHAVi dataset (8 cameras)[69]

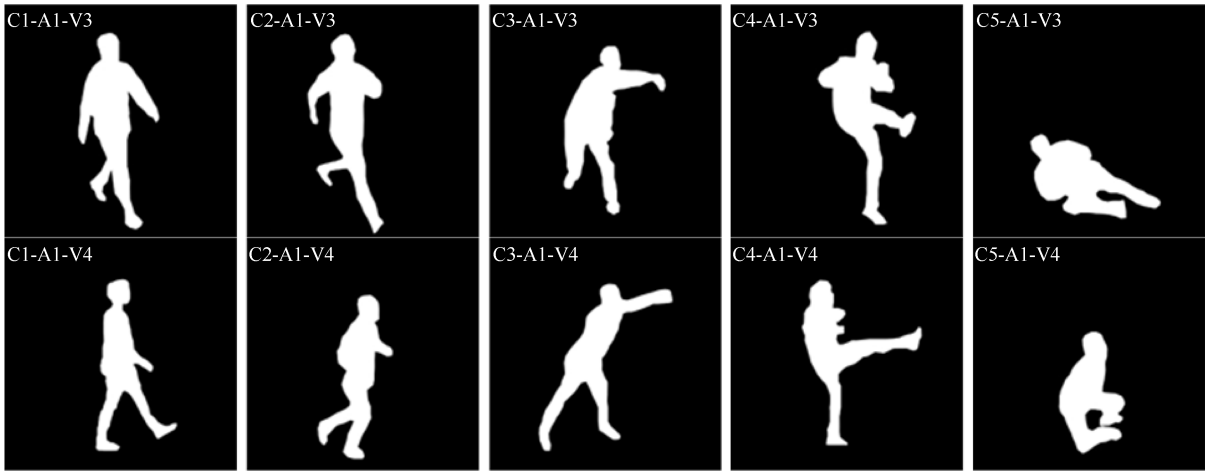


图 13 MuHAVi-Mas 数据集的 2 个视角剪影示例图[69]

Fig. 13 Sample silhouette images of MuHAVi-MAS dataset (2 cameras)[69]

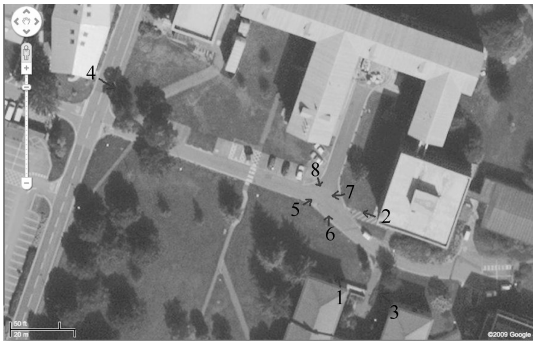


图 14 8 个摄像机位置和方向的平面图[70]

Fig. 14 Plan view showing the location and direction of the 8 cameras[70]

图 15 PETS 2009 基准数据集示例图[70]

Fig. 15 Sample images of PETS 2009 benchmark dataset[70]

集采用安装在车辆 4 个角落上的 4 个非重叠的视觉摄像机, 覆盖面积约 100 米 × 30 米, 如图 16 所示.

该数据集共包含 22 个视频, 其分辨率为 1 280 × 960, 其目的是检测和理解在停放的车辆周围的人类行为. 该数据集涉及视频理解的三个层次内容的挑战: 1) 低级视频分析, 即目标检测和跟踪; 2) 中级视频分析, 即简单事件检测, 涉及个体行为识别; 3) 高级视频分析, 即复杂事件检测, 涉及群体行为和交互行为识别. 该数据集主要侧重于区分正常、异常和威胁行为. 对威胁行为分为三个等级: 异常行为、潜在犯罪行为和犯罪行为, 其示例如图 17 所示. ARENA 数据集由于其复杂性, 在 PETS 2015[89] 和 PETS 2016[90] 中继续作为基准数据集之一使用.

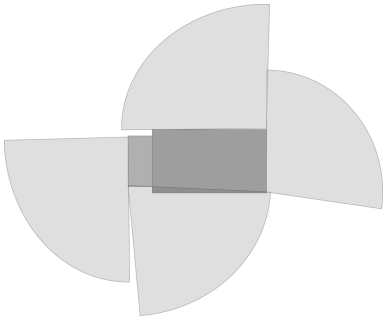


图 16 卡车车载摄像头位置及覆盖范围[71]

Fig. 16 The on-board camera configuration and coverage[71]

(a) 异常行为 (a) Abnormal behavior (b) 潜在犯罪行为 (b) Potentially criminal (c) 犯罪行为 (c) Criminal behavior

图 17 停放车辆周围的三种不同行为[91]

Fig. 17 Three different kinds of behavior recorded around a parked vehicle[91]

多视角数据集具有同一位置不同视角的信息, 有利于研究者进行视角无关的行为识别研究. 目前, 对多视角数据集, 研究者大都通过提取不同的特征 (如 STIP、Cuboid、MoSIFT、Hog3D、CMS 等), 采用不同的方法 (如字典学习、迁移学习、多任务学习等) 进行研究.

多视角数据集中各数据集的最新识别率、研究方法、评价方案等信息如表 3 所示.

表 3 多视角数据集的最新研究成果概览表

Table 3 Summary of state-of-the-art research results on multi-view datasets

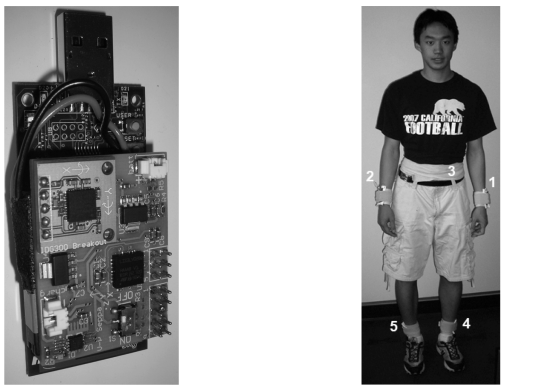
数据集名称	最新识别率	年份	研究方法	评价方案	备注
IXMAS	91.6 %[72]	2015	epipolar geometry	not mentioned	5 种行为
(单视角)	92.7 %[73]	2016	multi-view transition HMM	LOSOCV	11 种行为
IXMAS	95.54 %[75]	2014	MMM-SVM	Tr: one camera′s data	11 种行为; 5 个视角
(多视角)	95.3 %[74]	2016	Cuboid + supervised dictionary learning	LOAOCV; CV	11 种行为; 5 个视角
	95.1 %[74]	2016	STIP + supervised dictionary learning	LOAOCV; CV	11 种行为; 5 个视角
	95.54 %[75]	2014	MMM-SVM	Tr: one camera′s data Ts: LOSOCV	11 种行为; 4 个视角
	94.7 %[40]	2017	HC-MTL + L/S Reg	LOSOCV	11 种行为; 4 个视角
	93.7 %[92]	2017	eLR ConvNet(TCNN)	LOSOCV	12 种行为; 5 个视角
	85.8 %[46]	2016	SDEG + $\Re$ transform	LOOCV	13 种行为; 5 个视角
MuHAVi	97.48 %[83]	2012	Visual + Correlation (LKSSVM)	LOOCV	4 个视角
	92.1 %[82]	2014	sectorial extreme points (HMM)	LOSOCV	4 个视角
	91.6 %[84]	2016	CMS + multilayer descriptor (Multiclass K-NN)	LOOCV	8 个视角
MuHAVi-14	98.53 %[86]	2014	Pose dictionary learning + maxpooling	LOOCV	
	98.5 %[85]	2013	radial summary feature + Feature Subset Selection	leave-one-sequence-out	
	95.6 %[84]	2016	CMS + multilayer descriptor(Multiclass K-NN)	LOOCV	
	94.12 %[88]	2014	CMS (K-NN)	multi-training	
MuHAVi-8	100 %[84]	2016	CMS + multilayer descriptor (Multiclass K-NN)	LOOCV	
	100 %[88]	2014	CMS (K-NN)	multi-training	
	100 %[87]	2014	radial silhouette-based feature (multiview learing)	leave-one-sequence-out	
	100 %[85]	2013	radial summary feature + Feature Subset Selection	leave-one-sequence-out LOSOCV	

*CV: cross-view

4 特殊数据集

为了更好地研究人体运动过程中的运动规律, 采用特殊技术捕获动作数据, 为人体行为识别提供有利信息, 比如利用运动传感器、惯性传感器、红外摄像头、Kinect 相机等捕获运动信息、深度信息、人体骨架信息等. 常见的数据集有: WARD[93]、CMU Motion Capture[94]、MSR Action 3D[95]、MSR Daily Activity 3D[96]、UCF Kinect[97] 等.

WARD (Wearable Action Recognition Database)[93] 人体日常行为数据库来自美国加州大学伯克利分校部分支持的项目. 该数据库将无线运动传感器 (如图 18 (a) 所示) 放置于人体腰部、左右手腕和左右脚踝 5 个部位上 (如图 18 (b) 所示), 构成一个身体传感器系统. 其中, 每个传感器单元包括一个三轴加速度计和一个双轴陀螺仪, 数据采样频率为 20 Hz. 该数据集的早期规模较小, 利用 8 个无线运动传感器, 仅包含 3 个受试者 12 种行为类别的 626 个行为样本[98]. 目前该数据库包括年龄在 19 岁到 75 岁之间的 20 个受试者 (13 名男性和 7 名女性) 在自然状态下执行的 13 种行为, 共计 1 300 个行为样本 (现在实际有 1 298 个). 该数据库的 13 种行为类别为: 站、坐、躺、向前走、逆时针走、顺时针走、向左转、向右转、上楼、下楼、慢跑、跳和推轮椅, 每种行为重复执行 5 次.



(a) 无线运动传感器 (b) 身体传感器系统

(a) A wireless motion sensor (b) A body sensor system

图 18 WARD 数据库示例图[93]

Fig. 18 Sample images of WARD database[93]

该数据库除提供相对稳定的公开定量比较平台外, 还有望引导未来分布式模式识别领域的创新算法的发展. 目前, Guo 等[99] 首先对每个传感器节点的特征利用广义判别分析 (GDA) 进行降维, 然后采用多级关联向量机 (RVM) 获得个体分类, 最后利用传感器节点的异构和互补信息在决策层进行融合, 达到 98.78 % 的识别率. 而 Guo 等[100] 提出一种新的特征提取方法鲁棒线性判别分析 (RLDA), 通过主成分分析 (PCA) 降维后重新估计类内散射矩阵的特征值而获得新的投影矩阵, 达到 99.02 % 的识别率.

CMU Motion Capture (Mocap)[94] 数据集是由美国卡内基梅隆大学的图形实验室发布的. 该数据集采用 8 个红外摄像头, 提供带有 41 个标记关节点的信息, 可以精确估计人体骨架结构信息. 该数据集的运动捕获数据包括 6 个类别和 23 个亚类的 2 605 个实验. 每个实验包含一个或多个行为类别, 提供低分辨率的 RGB 视频和 3 种格式的关节点数据: tvd、c3d 和 amc. 6 个大类分别为: 人类交互、与环境交互、人体移动、体育活动和运动、情况和情景、测试运动, 其部分示例如图 19 所示.

图 19 CMU Mocap 数据集示例图

Fig. 19 Sample images of CMU Mocap dataset dataset

虽然 CMU Mocap 数据集随机采样执行动作, 其类内、类间的差异巨大, 但是由于提供的参数数据能够构建完整的 3D 模型, 吸引众多研究者的关注. 目前, 研究者从该数据集中选取不同类别进行研究. 对 5 种常见行为 (走、跳、跑、爬和高尔夫挥杆), Jia 等[101] 提出用于描述 3D 非共面点的投影不变量, 即特征数 (CN). 对运动轨迹, 利用时间序列的人体单个关节点计算视角无关的时间特征数 (TCN), 可用有限的关节点表征动作; 而在单帧的空间域, 计算 5 个关节点的空间特征数 (SCN), 其与时间特征具有互补性. 利用近邻分类器 (1-NN), 采用时间特征数达到 94.8 % 的识别率, 采用空间特征数达到接近 100 % 的识别率. Aghbari 等[102] 提出一种贪心算法 DisCoSet, 通过递增寻找一个最小的局部特征对比集, 不需要离散化就可以最大限度地区分一个类, 在选取的 12 种行为上达到 98.6 % 的识别率. 而 Kadu 等[103] 提出基于树型矢量量化 (TSVQ) 的多分辨率字符串表示方案将人体姿态的时间序列转换为码字序列, 并利用码字匹配考虑姿态的时间变化, 采用基于姿态直方图的支持向量机 (SVM) 进行分类, 在选取的 30 种行为上达到 99.6 % 的识别率.

利用 Microsoft Kinect 相机 (如图 20 所示) 采集的深度数据可获得较为精准的人体关节点骨架序列. 微软剑桥研究院 (Microsoft Research Cambridge, MSR) 先后发布了 MSR Action 3D[95] 和 MSR Daily Activity 3D[96], 美国中佛罗里达大学发布了 UCF Kinect[97]. 近几年, 陆续出现了综合利用 Kinect 和其他信息构建的多模态数据集, 如 N-UCLA Multiview Action3D[104]、UTD-MHAD[105] 等. 这些数据集都是基于 Kinect v1 (如图 20 (a) 所示) 构建的. 而随着 Kinect v2 (如图 20 (b) 所示) 的发布, 新加坡南洋理工大学的 Shahroudy 等[106] 利用其特点构建了包含 4 种模态的大型数据集 NTU RGB+D.

(a) 微软 Kinect v1
(a) Microsoft Kinect v1

(b) 微软 Kinect v2
(b) Microsoft Kinect v2

图 20 Microsoft Kinect 相机示例图

Fig. 20 Sample images of Microsoft Kinect camera

MSR Action 3D[95] 数据集提供 20 个关节点的三维坐标数据、深度图像与 RGB 图像, 包含 20 种行为类别, 每种行为由 10 个受试者重复执行 2 ∼ 3 次, 总共 567 个样本. 该数据集的 20 种行为类别为: 高挥手、水平挥手、锤、手抓、打拳、高抛、画叉、画勾、画圆、拍手、双手挥、侧边拳击、弯曲、向前踢、侧踢、慢跑、网球挥拍、网球发球、高尔夫挥杆、捡起扔 (对应标记为 a01 ∼ a20), 其中网球发球的深度序列图如图 21 所示.

图 21 MSR Action 3D 数据集的深度序列图[95]

Fig. 21 The sequences of depth maps of MSR Action 3D dataset[95]

MSR Action 3D 数据集的视频序列为无背景的纯人体运动, 但由于相似的动作以及关节位置噪声, 仍然非常具有挑战性. 为了减少测试的计算复杂度, 依据行为的复杂程度将数据集划分为 3 个子集: AS_1、AS_2 和 AS_3 (如表 4 所示). 其中每个子集包含 8 种行为类别, 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; 而子集 AS_3 中的动作复杂度最高.

表 4　MSR Action 3D 数据集的子集

Table 4　The subsets of MSR Action 3D dataset

数据子集	包含行为类别
AS_1	a02、a03、a05、a06、a10、a13、a18、a20
AS_2	a01、a04、a07、a08、a09、a11、a14、a12
AS_3	a06、a14、a15、a16、a17、a18、a19、a20

该数据集被研究者广泛研究, 已成为 3D 行为识别的典型基准数据集. 研究者大都采用划分 3 个子集和交叉受试者的方式进行验证. 在划分 3 个子集的情况下, Luo 等[107] 提出基于组稀疏和几何约束的字典学习 (DL-GSGC) 算法, 利用时间金字塔匹配 (TPM), 在利用 1/3 样本和 2/3 样本进行训练时均达到 98.9 % 的识别率. 而 Chen 等[108] 采用来自三个投影视图的深度运动图 (DMM) 捕捉运动线索, 同时使用局部二值模式 (LBP) 获得紧凑特征表征, 利用特征级和决策级两种融合方式, 在利用 2/3 样本进行训练时达到 100 % 的识别率. 在交叉受试者的情况下, Chen 等[109] 提出一个有效利用 3D 深度数据进行识别的框架 TriViews, 通过对每个投影视图的 5 个不同特征 (STIP、DT-Shape、DT-MBH、ST-Shape 和 ST-MBH) 选取最佳三个特征基于概率融合方法 (PFA) 进行融合, 达到 98.2 % 的识别率. 而澳大利亚卧龙岗大学高级多媒体研究实验室的 Wang 等[110] 提出利用分层深度运动图 (HDMM) 和 3 通道深度卷积神经网络 (3ConvNets) 的框架对深度图序列进行识别, 达到 100 % 的识别率.

MSR Daily Activity 3D[96] 是由 Kinect 设备捕获的日常活动的数据集. 该数据集由 10 个受试者执行, 包含 16 种类别的 320 个样本. 该数据集的 16 种日常行为类别为: 喝、吃、读书、打手机、写字、用笔记本电脑、用吸尘器、欢呼、静坐、扔纸、玩游戏、躺沙发、走、弹吉他、站起来、坐下, 其示例如图 22 所示. 其中, 每种行为由受试者以站姿或坐姿分别执行 2 次, 因此严格说来, 该数据集的行为类别分为 17 种, 因为静坐在执行时分别执行了两类行为: 静坐和站. 该数据集在具有背景物体的真实环境拍摄, 并且受试者距离相机的位置不固定; 大部分样本涉及到人与物体的交互行为; 有些行为包含身体的细节运动; 捕获的 3D 关节点坐标受噪声污染严重. 因此, 该数据集比 MSR Action3D[95] 数据集更具挑战性. 截至目前, Zhang 等[111] 通过深度梯度信息和骨架关节点距离来提取粗 Depth-Skeleton (DS) 特征, 并利用稀疏编码和最大池化进行细化, 采用随机决策森林 (RDF) 进行分类达到 97.5 % 的识别率; 而 Shahroudy 等[112] 考虑 RGB 信息和深度信息的互补性, 提出一种基于共享特性特征分解网络的深度自动编码器, 将输入的多模态信号分离成一个分层结构, 利用结构化稀疏学习机 (SSLM) 同样获得 97.5 % 的识别率.

UCF Kinect[97] 数据集使用微软 Kinect 传感

图 22　MSR Daily Activity 3D 数据集示例图

Fig. 22　Sample images of MSR Daily Activity 3D dataset

器和 OpenNI 平台估计骨架, 包含 16 个受试者 (13 个男性和 3 个女性), 年龄介于 20 岁到 35 岁之间, 共计 1 280 个行为样本. 该数据集的 16 种行为类别为: 平衡、向上爬、爬梯子、躲避、单脚跳、跳跃、飞跃、跑、踢、打拳、向左扭、向右扭、向前走、后退、向左速移和向右速移, 其中每种行为由每个受试者重复执行 5 次. 而且在每帧中, 包含 15 个关节点的三维坐标及方向数据, 部分骨架示例如图 23 所示.

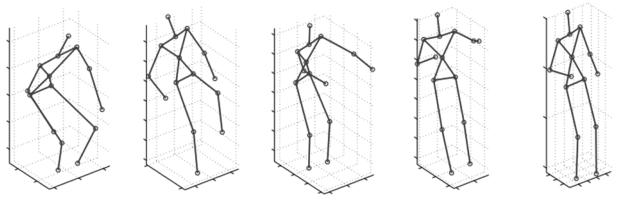


图 23 UCF Kinect 数据集的骨架示例图[97]

Fig. 23 Sample skeleton images of UCF Kinect dataset[97]

该数据集在收集每个行为数据时, 要求受试者以一个放松的姿势站立, 双手自然垂于身体两侧, 因此, 可以更真实地估计各种行为的等待时间. 该数据集具有不同的视点, 且相同行为具有类内差异. Kerola 等[113] 利用深度图序列, 基于骨架和关键点分别利用光谱图小波变换 (SGWT) 和金字塔池化计算相应的光谱图序列 (SGS) 描述符, 再通过 SVM 训练并使用晚融合策略达到 98.8 % 的识别率. 而 Beh 等[114] 为在单位超球面空间对手势轨迹建模, 将 MvMM 概率密度函数并入 HMM, 同时利用 L_2 正则化达到 98.9 % 的识别率.

N-UCLA Multiview Action3D[104] 数据集由美国西北大学和加州大学洛杉矶分校联合构建. 该数据集将深度、骨架和多视角数据融合在一起, 旨在捕获人类从多个摄像机角度执行的日常行为. 该数据集由 3 个 Kinect 相机从三个视角同时捕获, 包含 10 个受试者执行 10 种日常行为的 1 493 个行为样本 (现在实际有 1 475 个). 10 种日常行为是: 用一只手捡 (Pick up with one hand)、用两只手捡 (Pick up with two hands)、丢垃圾 (Drop trash)、走动 (Walk around)、坐下 (Sit down)、站起来 (Stand up)、穿衣 (Donning)、脱衣 (Doffing)、投掷 (Throw) 和搬运 (Carry), 其示例如图 24 所示.

该数据集的若干行为包括与对象的交互, 如丢垃圾和搬运; 每个动作都是从不同的视角捕获的, 其视角分布如图 25 所示; 有些行为非常相似, 如用一只手捡和用两只手捡; 有些动作很容易误判, 如将丢垃圾误认为是走动. 因此, 该数据集非常具有挑战性. Kerola 等[113] 利用骨架和关键点构建的 SGS 取得 90.8 % 的识别率. 而 Liu 等[115] 针对时空骨架序列的有效表征问题提出一种增强骨架可视化方法, 通过基于序列的视角无关变换将骨架序列可视化为一系列彩色图像, 并对彩色图像利用视觉和运动增强方法进行局部增强, 然后利用 CNN 模型在决策级融合, 达到 92.61 % 的识别率.

UTD-MHAD[105] 数据集是由德克萨斯大学达拉斯分校的机构审查委员会 (IRB) 发布的多模态人体行为识别数据集. 该数据集由 Kinect 相机和可穿戴惯性传感器 (如图 26 (a) 所示) 同时来捕获 4 种模式的数据: RGB 视频、深度视频、骨架关节点位置和惯性传感器信号, 其左臂向右滑行为的多模态数据示例如图 27 所示. 这 4 种模式的数据被记录在 3 个通道, 其中深度视频和 20 个骨架关节点位置信息被同时捕获在一个通道. 该数据集包含 27 种行为, 由 8 名受试者 (4 名男性和 4 名女性) 重复执行 4 次, 共计 861 个样本 (去掉了 3 个损坏样本).

该数据集的 27 种行为可以分为 4 大类: 1) 体育运动: 篮球投篮、保龄球、正面拳击、棒球挥杆、网球正手挥拍、网球发球; 2) 手势: 左臂向左滑、左臂向右滑、画 X、顺时针画圆、逆时针画圆、画三角形; 3) 日常活动: 挥手、两手前拍、扔、交叉双臂、双

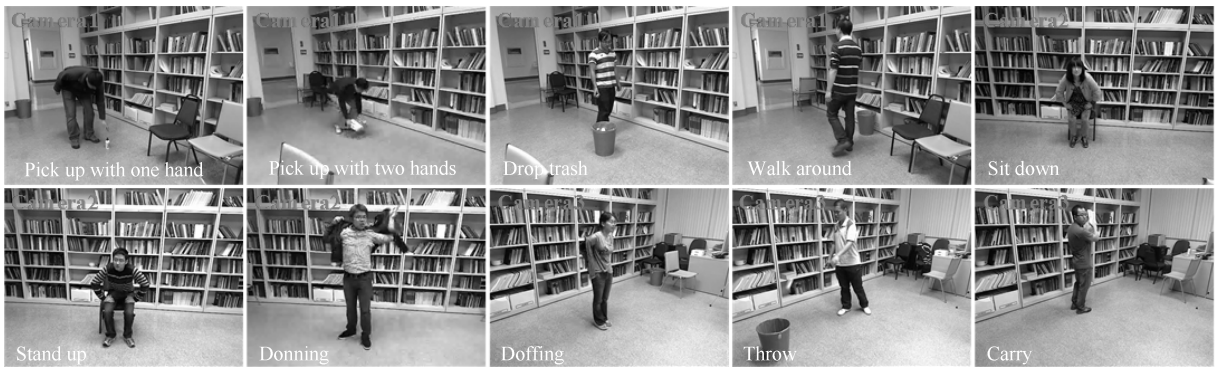


图 24 N-UCLA Multiview Action3D 数据集示例图

Fig. 24 Sample images of N-UCLA Multiview Action3D dataset

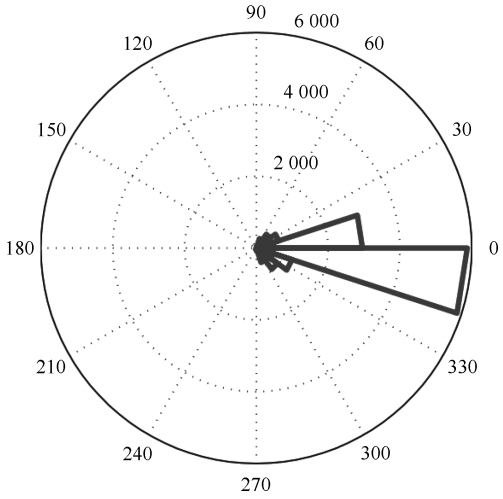
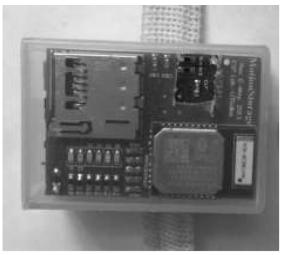


图 25　Multiview Action3D 的视角分布[104]

Fig. 25　The view distribution of Multiview Action3D dataset[104]

手推、敲门、抓物、捡起扔、慢跑、走、站起来、坐下; 4) 训练练习: 双臂二头肌弯曲、左脚向前弓步、伸臂蹲. 在采集数据集时, 可穿戴惯性传感器位于右



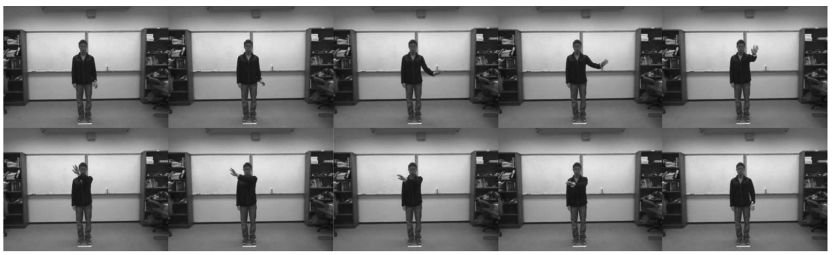
(a) 可穿戴惯性传感器
(a) A wearable inertial sensor

(b) 右手腕
(b) On right wrist

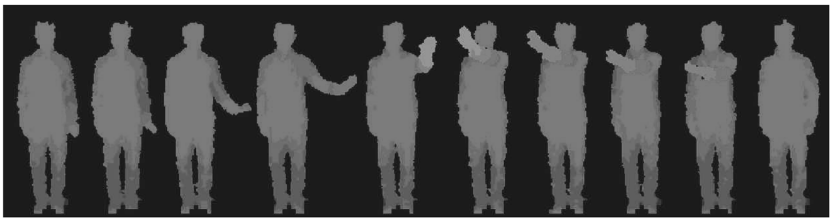
(b) 右大腿
(b) On right thigh

图 26　可穿戴惯性传感器及其位置示例图[105]

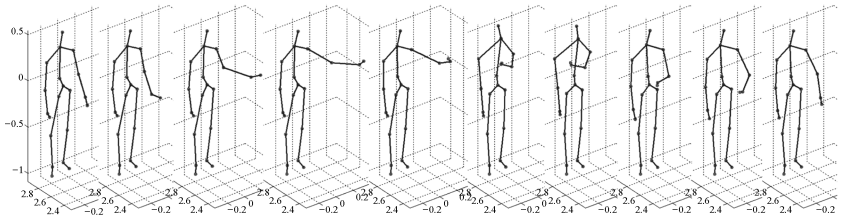
Fig. 26　Sample images of the wearable inertial sensor and its placements[105]



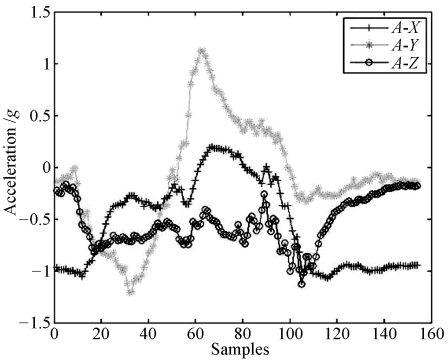
(a) RGB 图像
(a) The RGB images



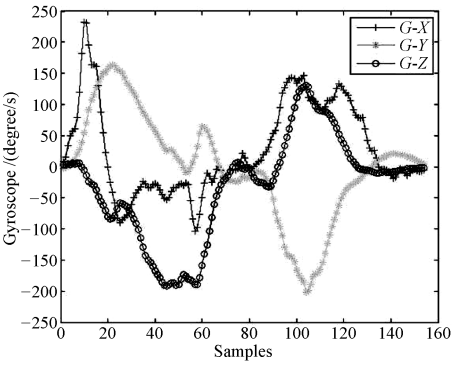
(b) 深度图像
(b) The depth images



(c) 骨架关节点图像
(c) The skeleton joint images



(d) 惯性传感器数据 (陀速度信号)
(d) The inertial sensor data (acceleration signals)



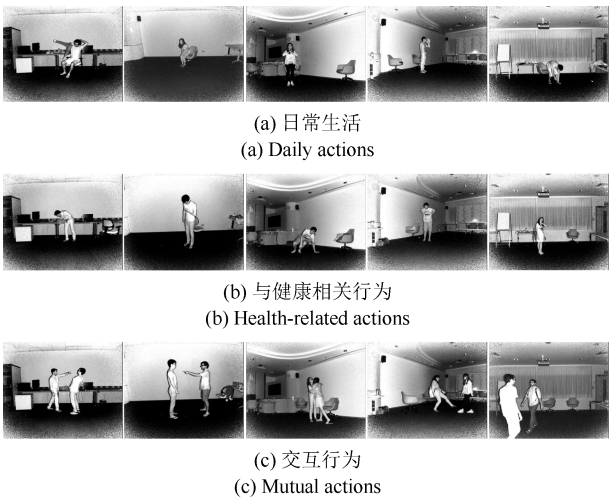
(e) 惯性传感器数据 (陀螺仪信号)
(e) The inertial sensor data (gyroscope signals)

图 27　左臂向右滑行为的多模态数据示例图

Fig. 27　Sample images of the multimodality data corresponding to the action left arm swipe to the right

手腕 (21 种行为) 或右大腿 (6 种行为), 如图 26 (b) 和 (c) 所示 (实际测试时位于左手腕或左大腿). 由于受试者的差异, 并且行为以自然方式在不同的速度下执行, 因此该数据集具有较大的类内变化, 非常具有挑战性. 目前, Li 等[116] 通过关节距离图 (JDM) 将 3D 骨架序列转化为 4 个二维彩色图像, 同时采用 4 个 CNN 分别学习判别特征, 通过晚融合获得 88.1 % 的识别率. 而 Bulbul 等[117] 从整个视频序列生成三个 DMM, 然后利用 DMM 获得三个判别特征: 基于轮廓的方向梯度直方图 (CT-HOG)、局部二值模式 (LBP) 和边缘方向直方图 (EOH), 最后采用决策级融合达到 88.4 % 的识别率.

NTU RGB+D[106] 数据集是由新加坡南洋理工大学的博云搜索实验室 (Rapid-Rich Object Search, ROSE) 于 2016 年发布的最新的多视角深度信息数据集. 利用 Kinect v2 的高分辨率和新的主动式红外检测, 构建了包含 4 种模态的大型数据集: RGB 视频、深度视频、骨架关节点位置和红外视频. 该数据集由年龄介于 10 岁到 35 岁之间的 40 个受试者执行 60 种行为, 共计 56 880 个行为样本, 4 种模态数据共计 1.3 TB. 该数据集也是多视角数据集, 由 3 个 Kinect v2 相机从三个角度的 17 种不同高度和距离同时捕获, 共计 80 个视角. 该数据集的行为类别分成三类: 1) 40 种日常行为; 2) 9 种与健康相关行为; 3) 11 种交互行为, 其红外视频的部分示例如图 28 所示. 该数据集利用 Kinect v2 获得具有 25 个骨架关节点的信息, 其分布示意图如图 29 所示.



(a) 日常生活
(a) Daily actions

(b) 与健康相关行为
(b) Health-related actions

(c) 交互行为
(c) Mutual actions

图 28 NTU RGB+D 数据集的红外示例图

Fig. 28 Sample infrared images of NTU RGB+D dataset

NTU RGB+D 数据集不仅包含复杂的行为类型和多模态的数据信息, 而且数据量非常大, 具有很大挑战性. 该数据集在 2016 年的 CVPR 会议上一经提出, 立即引起研究者的关注. 针对该数据集的特点, 研究者大都采用 Shahroudy 等[106] 提出的两种测试验证方式 (交叉受试者验证和交叉视角验证). 交叉受试者验证的训练集包含 20 个受试者共计 40 320 个样本; 测试集包含 20 个受试者共计 16 560 个样本. 而交叉视角验证的训练集包含相机 2 和 3 的视频, 共计 37 920 个样本; 测试集包含相机 1 的视频, 共计 18 960 个样本. Wang 等[118] 提出了一种简单有效的表征 3D 骨架序列时空信息的方法, 通过关节轨迹图 (JTM) 将 3D 骨架序列转化为三个二维彩色图像, 同时采用三个 CNN 分别学习判别特征, 并通过多分数层融合 (MSF) 提高识别准确度. 该方法在交叉受试者的方式下, 达到 76.32 % 的识别率; 而在交叉视角的方式下, 达到 81.08 % 的识别率. Li 等[116] 提出的利用关节距离图 (JDM) 方法在交叉受试者的方式下, 达到 76.2 % 的识别率; 而在交叉视角的方式下, 达到 82.3 % 的识别率. 由此可以看出, 关节轨迹图 (JTM) 和关节距离图 (JDM) 各有优势, 二者的关系有待进一步探索.

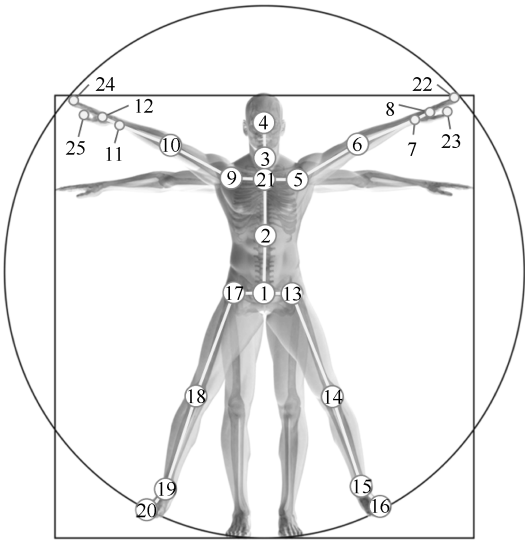


图 29 25 个骨架点示意图[106]

Fig. 29 Configuration of 25 body joints[106]

特殊数据集, 尤其是 RGB-D 数据集, 由于其提供的多模态信息的互补性而受到研究者的广泛关注. 研究者利用深度、骨架等信息, 通过深度图序列、3D 骨架序列等提取不同判别特征来提高识别率. Li 等[116] 和 Wang 等[118] 将 3D 骨架序列进行转换后, 利用 CNN 学习判别特征的新思路值得借鉴. 而 Zhang 等[14] 提出在 RGB-D 数据集中采用交叉数据集验证方式增强数据集鲁棒性和实用性的建议有待进一步研究. 此外, 随着红外视频数据集的发展, 红外信息具有的避免光照、阴影、遮挡等因素影响的特性也将受到研究者的关注.

近年来, 随着对老人、孩子等特殊群体安全

及监护的需求, 相继出现了包含跌倒行为在内的日常行为数据集, 如 UR Fall Detection Dataset (URFD)[119]、TST Fall Detection v1[120]、TST Fall Detection v2[121] 等, 也给人体行为识别的研究提出了新要求.

特殊数据集中各数据集的最新识别率、研究方法、评价方案等信息如表 5 所示.

5 公开数据集比较

本文对上述介绍的 4 类人体行为数据库/集, 从公开年份、行为类别、行为人数、视频总数、每类视频数、分辨率等方面进行了的详细比较, 其信息如表 6 和表 7 所示. 从表中可以看出, 特殊数据集的行为类别和规模相对于真实场景数据库来说较少. 这与特殊数据库需要利用专门的设备来捕获有直接的关系. 另外, 根据这 4 类数据集的场景、内容、视角、应用领域等信息, 对各数据集按不同特征进行分类对比, 具体内容如表 8 所示.

由于篇幅所限, 本文中仅介绍了相对应用较多的公开数据集. 还有一些数据集信息参见表 6～8.

表 5 特殊数据集的最新研究成果概览表

Table 5 Summary of state-of-the-art research results on special datasets

数据集名称	最新识别率	年份	研究方法	评价方案	备注
WARD	99.02 %[100]	2015	PCA+RLDA (SVM)	CS: Tr: 15; Te: 5	
	98.78 %[99]	2012	GDA+RVM+WLOGP	3-fold cross-validation	
	97.5 %[122]	2017	FDA (SVM)	20-fold cross-validation	10 种行为
	近 100 %[101]	2016	SCN (1-NN)	CS	5 种行为 200 个样本
CMU Mocap	98.27 %[123]	2010	HGPLVM	3-fold cross-validation	5 种行为
	98.13 %[124]	2014	3D joint position features+Actionlet Ensemble	not mentioned	5 种行为
	98.6 %[102]	2015	DisCoSet (SVM)	All	12 种行为 164 个样本
	99.6 %[103]	2014	TSVQ (Pose-Histogram SVM)	5-fold cross-validation	30 种行为 278 个样本
MSR Action 3D (AS_1、AS_2 和 AS_3)	100 %[108]	2015	DMM-LBP-FF/DMM-LBP-DF	Tr: 2/3; Te: 1/3	
	98.9 %[107]	2013	DL-GSGC	Tr: 2/3; Te: 1/3	
	98.9 %[107]	2013	DL-GSGC	Tr: 1/3; Te: 2/3	
	98.7 %[108]	2015	DMM-LBP-FF	Tr: 1/3; Te: 2/3	
	96.7 %[107]	2013	DL-GSGC	CS	
	96.1 %[125]	2016	3D skeleton+two-level hierarchical framework	CS	
	96.0 %[111]	2017	Coarse DS+Sparse coding (RDF)	CS	
MSR Action 3D (cross-subject)	100 %[110]	2015	HDMM+3ConvNets	Tr: 奇数; Te: 偶数	
	98.2 %[109]	2015	TriViews+ PFA	Tr: 奇数; Te: 偶数	
	98.2 %[126]	2015	Decision-Level Fusion (SUM Rule)	Tr: 2/3/5/7/9; Te: 1/4/6/8/10	
	96.7 %[107]	2013	DL-GSGC+TPM	Tr: 奇数; Te: 偶数	
MSR Daily Activity 3D	97.5 %[111]	2017	Coarse DS+Sparse coding (RDF)	not mentioned	
	97.5 %[112]	2016	DSSCA+SSLM	CS	
	95.0 %[107]	2013	DL-GSGC+TPM	CS	
UCF Kinect	98.9 %[114]	2014	MvMF-HMM+L_2-normalization	4-fold cross-validation	
	98.8 %[113]	2017	SGS($p_{\text{mean}}/p_{\max}$, skeleton-view-dep.)	4-fold cross-validation	
	98.7 %[127]	2013	motion-based grouping+adaptive weighting (hierarchical model)	2-fold cross-validation	

表 5 特殊数据集的最新研究成果概览表 (续)

Table 5 Summary of state-of-the-art research results on special datasets (Cont)

数据集名称	最新识别率	年份	研究方法	评价方案	备注
N-UCLA	92.61 %[115]	2017	Synthesized+Pre-trained (CNN)	CV	
Multiview Action 3D	90.8 %[113]	2017	SGS($p_{\max}$, skel.-view-inv.+keypoint)	CV	
	89.57 %[115]	2017	Synthesized Samples (CNN)	CV	
	81.6 %[104]	2014	MST-AOG	CS; LOOCV	
	79.3 %[104]	2014	MST-AOG	cross-environment	
UTD-MHAD	88.4 %[117]	2015	DMMs+CT-HOG+LBP+EOH	CS	
	88.1 %[116]	2017	JDM+MSF (CNN)	CS	
	87.9 %[118]	2016	JTM+MSF (CNN)	CS	
NTU RGB+D	76.32 %[118]	2016	JTM+MSF (CNN)	CS	
	76.2 %[116]	2017	JDM+MSF (CNN)	CS	
	62.93 %[106]	2016	2layer P-LSTM	CS	
	82.3 %[116]	2017	JDM+MSF (CNN)	CV	
	81.08 %[118]	2016	JTM+MSF (CNN)	CV	
	70.27 %[106]	2016	2 layer P-LSTM	CV	

表 6 通用、真实场景及多视角数据集信息表

Table 6 The information of general datasets, real scene datasets and multi-view datasets

类型	数据集名称	年份	行为类别	行为人数	视频数/类	视频总数/样本数	场景	视角	分辨率 (最高)	fps
通	KTH[19]	2004	6	25	99 ∼ 100	599/2 391	4	1	160×120	25
用	Weizmann[2]	2005	10	9	9 ∼ 10	93	1	1	180×144	25
	Hollywood[27]	2008	8	N/A	30 ∼ 129	475	N/A	N/A	544×240	25
	UCF Sports[28]	2008	10	N/A	6 ∼ 22	150	N/A	N/A	720×480	9
	UT-Tower[128]	2009	9	6	12	108	2	1	360×240	10
	Hollywood 2[29] (Actions)	2009	12	N/A	61 ∼ 278	2 517	N/A	N/A	720×528	25
	ADL[129]	2009	10	5	15	150	1	1	1 280×720	30
	UCF YouTube[30]	2009	11	N/A	116 ∼ 198	1 600	N/A	N/A	320×240	30
	Olympic Sports[31]	2010	16	N/A	21 ∼ 67	783	N/A	N/A	–	–
真	UT-Interaction[130]	2010	6	N/A	20	120	2	1	720×480	30
实	HMDB51[32]	2011	51	N/A	102 ∼ 548	6 766	N/A	N/A	424×240	30
场	CCV[131]	2011	20	N/A	224 ∼ 806	9 317	N/A	NA	–	–
景	UCF50[33]	2012	50	N/A	100 ∼ 197	6 681	N/A	N/A	320×240	25
	UCF101[34]	2012	101	N/A	100 ∼ 167	13 320	N/A	N/A	320×240	25
	MPII Cooking[132]	2012	65	12	–	44/5 609	1	1	1 624×1 224	29.4
	MPII Composites[133]	2012	60	22	–	212	1	1	1 624×1 224	29.4
	Sports-1M[35]	2014	487	N/A	1 000 ∼ 3 000	1 133 158	N/A	N/A	1 280×720	30
	Hollywood Extended[134]	2014	16	N/A	2 ∼ 11	937	N/A	N/A	720×528	25
	MPII Cooking 2[135]	2015	67	30	–	273/14 105	1	1	1 624×1 224	29.4
	ActivityNet[136]	2015	203	N/A	137(a)	27 801	N/A	N/A	1 280×720	30
多	IXMAS[68]	2006	13	12	180	180/2 340	1	5	390×291	23
视	i3DPost[137]	2009	12	8	64	768	1	8	1 920×1 080	25
角	MuHAVi[69]	2010	17	7	56	952	1	8	720×576	25
	MuHAVi-MAS[69]	2010	14	2	4 ∼ 16	136	1	2	720×576	25

*a: average; N/A: not applicable

表 7 特殊数据集信息表

Table 7 The information of special human activity recognition datasets

数据集名称	年份	行为类别	行为人数	视频数/类	视频总数/样本数	场景	视角	分辨率	fps	数据格式	骨架关节点
CMU Mocap[94]	2007	23 个亚类	N/A	1～96	2 605	N/A	N/A	320 × 240	30	MS	41
WARD[93]	2009	13	20	64～66	1 298	1	1	–	–	M	N/A
CMU-MMAC[138]	2009	5 大类	45	234～252	1 218	1	6	1 024×768 640×480	30 60	RDMA	N/A
MSR Action 3D[95]	2010	20	10	20～30	567	1	1	640×480 (R) 320×240 (D)	15	DS	20
RGBD-HuDaAct[139]	2011	12	30	–	1 189	1	1	640×480 (RD)	30	RD	N/A
UT Kinect[140]	2012	10	10	–	200	1	1	640×480 (R) 320×240 (D)	30	RDS	20
ACT4^2[141]	2012	14	24	–	6 844	1	4	640 × 480	30	RD	N/A
MSR Daily Activity 3D[96]	2012	16	10	20	320	1	1	640×480	30	RDS	20
UCF Kinect[97]	2013	16	16	80	1 280	1	1	–	–	S	15
Berkeley MHAD[142]	2013	11	12	54～55	659	1	4	640×480	30	RDMAIe	N/A
3D Action Pairs[143]	2013	12	10	30	360	1	1	640×480	30	RDS	20
Multiview RGB-D event[144]	2013	8	8	477 (a)	3 815	1	3	640×480	30	RDS	20
Online RGBD Action[145]	2014	7	24	48	336	1	1	–	–	RDS	20
URFD[119]	2014	5	5	6～60	100	4	2	640×240	30	RD	N/A
N-UCLA[104]	2014	10	10	140～173	1 475	1	3	640×480	12	RDS	20
TST Fall detection v1[120]	2014	2	4	10	20	1	1	320×240 (D)	30	D	N/A
UTD-MHAD[105]	2015	27	8	31～32	861	1	1	640×480	30	RDSIe	25
TST Fall detection v2[121]	2016	8	11	33	264	1	1	512×424 (D)	25	DSIe	25
NTU RGB+D[106]	2016	60	40	948	56 880	1	80	1 920×720 (R) 512×424 (D) 512×424 (If)	30	RDSIf	25

*R: RGB; D: Depht; S: Skeleton; M: Motion; A: Audio; If: Infrared; Ie: Inertrial

表 8 人体行为数据集分类信息表

Table 8 Human activity dataset classification according to different features

分类特征	子类	数据集
场景	室内	ADL、MPII Cooking、MPII Composites、MPII Cooking 2、IXMAS、i3DPost、MuHAVi、MuHAVi-MAS、CMU Mocap、WARD、CMU-MMAC、MSR Action 3D、RGBD-HuDaAct、UT Kinect、ACT4^2、MSR Daily Activity 3D、UCF Kinect、MHAD、3D Action Pairs、Multiview RGB-D event、Online RGBD Action、URFD、N-UCLA Multiview Action 3D、TST Fall detection dataset v1、UTD-MHAD、TST Fall detection dataset v2、NTU RGB+D
	室外	Weizmann、UT-Tower、UT-Interaction、PETS
	室内/室外	KTH、Hollywood、UCF Sports、Hollywood 2、UCF YouTube、Olympic Sports、HMDB51、CCV、UCF50、UCF101、Sports-1M、Hollywood Extended、ActivityNet、THUMOS
内容	日常活动	KTH、Weizmann、ADL、HMDB51、CCV、ActivityNet、IXMAS、i3DPost、MuHAVi、MuHAVi-MAS、CMU Mocap、WARD、MSR Action 3D、RGBD-HuDaAct、UT Kinect、ACT4^2、MSR Daily Activity 3D、RGBD-HuDaAct、UCF Kinect、MHAD、3D Action Pairs、Multiview RGB-D event、Online RGBD Action、URFD、N-UCLA Multiview Action 3D、TST Fall detection dataset v1、UTD-MHAD、TST Fall detection dataset v2、NTU RGB+D
	体育运动	UCF Sports、UCF YouTube、Olympic Sports、UCF50、UCF101、Sports-1M、THUMOS
	厨房活动	MPII Cooking、MPII Composites、MPII Cooking 2、CMU-MMAC
	电影	Hollywood、Hollywood 2、Hollywood Extended
	监控	UT-Tower、UT-Interaction、PETS

表 8 人体行为数据集分类信息表 (续)

Table 8 Human activity dataset classification according to different features (Cont)

分类特征	子类	数据集
视角	单视角	KTH、Weizmann、ADL、MPII Cooking、MPII Composites、MPII Cooking 2、MSR Action 3D、UT Kinect、MSR Daily Activity 3D、RGBD-HuDaAct、UCF Kinect、3D Action Pairs、Online RGBD Action、TST Fall detection dataset v1、UTD-MHAD、TST Fall detection dataset v2
	多视角	IXMAS、i3DPost、MuHAVi、MuHAVi-MAS、ACT4^2、MHAD、Multiview RGB-D event、URFD、N-UCLA Multiview Action 3D、NTU RGB+D、PETS
	俯瞰	UT-Tower、UT-Interaction、PETS
	其他	Hollywood、UCF Sports、Hollywood 2、UCF YouTube、Olympic Sports、HMDB51、CCV、UCF50、UCF101、Sports-1M、Hollywood Extended、ActivityNet、CMU Mocap、WARD、CMU-MMAC、THUMOS
相机	静止	KTH、Weizmann、UT-Tower、ADL、UT-Interaction、MPII Cooking、MPII Composites、MPII Cooking 2、IXMAS、i3DPost、MuHAVi、MuHAVi-MAS、CMU-MMAC、MSR Action 3D、RGBD-HuDaAct、UT Kinect、ACT4^2、MSR Daily Activity 3D、UCF Kinect、MHAD、3D Action Pairs、Multiview RGB-D event、Online RGBD Action、URFD、N-UCLA Multiview Action 3D、TST Fall detection dataset v1、UTD-MHAD、TST Fall detection dataset v2、NTU RGB+D、PETS
	移动	Hollywood、UCF Sports、Hollywood 2、UCF YouTube、Olympic Sports、HMDB51、CCV、UCF50、UCF101、Sports-1M、Hollywood Extended、ActivityNet、CMU Mocap、THUMOS
应用	行为识别	KTH、Weizmann、Hollywood、UCF Sports、UT-Tower、Hollywood 2、ADL、UCF YouTube、Olympic Sports、UT-Interaction、HMDB51、CCV、UCF50、UCF101、MPII Cooking、MPII Composites、Sports-1M、Hollywood Extended、ActivityNet、MPII Cooking 2、IXMAS、i3DPost、MuHAVi、MuHAVi-MAS、CMU Mocap、WARD、CMU-MMAC、MSR Action 3D、RGBD-HuDaAct、UT Kinect、ACT4^2、MSR Daily Activity 3D、UCF Kinect、MHAD、3D Action Pairs、Multiview RGB-D event、Online RGBD Action、N-UCLA Multiview Action 3D、UTD-MHAD、TST Fall detection dataset v2、NTU RGB+D、PETS、THUMOS
领域	检测/跟踪	KTH、Weizmann、UCF Sports、Olympic Sports、UT-Interaction、ADL、UCF YouTube、ACT4^2、URFD、TST Fall detection dataset v1、TST Fall detection dataset v2、PETS、UCF50、UCF101、MPII Cooking、MPII Composites、MPII Cooking 2
	其他	KTH、Weizmann、UCF YouTube、UT-Tower、UCF50、ActivityNet、MPII Cooking、MPII Composites、MPII Cooking 2、Multiview RGB-D event

6 总结与展望

总体而言, 早期的公开数据集相机固定、行为类别较少、背景较简单. 而近几年发布的人体行为识别公开数据集有如下几个趋势:

1) 行为类别和数量越来越多. 随着科技的发展和设备的进步, 发布的公开数据集的行为类别从最初 KTH 的 6 种行为类别发展到 Sports-1M 的 487 种行为类别. 而视频的数量从 100 个左右发展到 1M. 近期, Google 又公布了一个大型视频数据集 YouTube-8M[146]. 该数据集是目前最大的视频数据集, 包含 800 万个 YouTube 视频共计 4 800 个类别, 并带有视频标注. 而其中与人相关的视频只是其中的一小部分, 大约有 8 000 个. 虽然如此, 但可以肯定, 人体行为识别公开数据集的规模会越来越大, 行为类别的数量会越来越多.

2) 行为越来越复杂. 公开数据集的人体行为从走、跑、跳等简单的行为发展到涉及人与人交互、人与物交互、异常行为、群体行为等复杂行为. 对异常行为、交互行为、群体行为等复杂行为的识别, 逐渐成为研究者关注的热点, 并将为以后公共场所的安全防范提供有力的保障.

3) 场景越来越复杂. 数据集的视频从简单场景到复杂场景, 并伴有遮挡、光照等噪声影响, 给人体行为识别的研究带来进一步的挑战. 因此, 如何降低噪声对识别效果的影响是人体行为识别未来的研究方向之一.

4) 多视角化. 较早的公开数据集相机基本固定, 几乎没有视角变化. 近几年的数据集出现了相机运动和视角变化. 而相机在不同视角下, 人、物和场景的大小、方向和形状都会发生变化, 这给行为识别的研究提出了新要求. 在行为识别中, 多视角的研究具有一定的优势, 通过视角变化对人体行为进行二维或三维建模, 利用相同点在模型不同位置的匹配和

分析实现不同视角下人体行为特征的表征. 因此, 视角无关的行为识别研究也是人体行为识别未来的研究方向之一.

5) 多模态化. 随着各式新型传感器和设备的发展, 相继出现了包含 RGB 视频、深度信息、骨架信息、红外信息等多模态信息的数据集. 不同模态数据之间存在较强相关性, 利用人体行为语义信息和互补性信息, 从多模态的低层特征学习到高层语义特征来进行人体行为识别, 这也将成为未来的研究方向.

总之, 人体行为识别公开数据集越来越接近于不受控的自然状态下的情形, 给研究者在保持算法鲁棒性的同时, 提高行为识别准确率带来更大的难度. 而随着深度学习在目标检测、分类等领域的应用, 其强大的数据表达能力, 必将为提高行为识别的性能开辟一个新的研究方向.

References

1 Hu W M, Tan T N, Wang L, Maybank S. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2004, **34**(3): 334−352

2 Kim I S, Choi H S, Yi K M, Choi J Y, Kong S G. Intelligent visual surveillance-a survey. *International Journal of Control, Automation and Systems*, 2010, **8**(5): 926−939

3 Huang Kai-Qi, Chen Xiao-Tang, Kang Yun-Feng, Tan Tie-Niu. Intelligent visual surveillance: a review. *Chinese Journal of Computers*, 2015, **38**(6): 1093−1118
(黄凯奇, 陈晓棠, 康运锋, 谭铁牛. 智能视频监控技术综述. 计算机学报, 2015, **38**(6): 1093−1118)

4 Dix A. *Human-Computer Interaction*. Berlin: Springer-Verlag, 2009. 1327−1331

5 Myers B A. A brief history of human-computer interaction technology. *Interactions*, 1998, **5**(2): 44−54

6 Rautaray S S, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 2015, **43**(1): 1−54

7 Park S H, Won S H, Lee J B, Kim S W. Smart home-digitally engineered domestic life. *Personal and Ubiquitous Computing*, 2003, **7**(3−4): 189−196

8 Jeong K-A, Salvendy G, Proctor R W. Smart home design and operation preferences of Americans and Koreans. *Ergonomics*, 2010, **53**(5): 636−660

9 Komninos N, Philippou E, Pitsillides A. Survey in smart grid and smart home security: Issues, challenges and countermeasures. *IEEE Communications Surveys & Tutorials*, 2014, **16**(4): 1933−1954

10 Suma E A, Krum D M, Lange B, Koenig S, Rizzo A, Bolas M. Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit. *Computers & Graphics*, 2013, **37**(3): 193−201

11 Zelnik-Manor L, Irani M. Event-based analysis of video. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Kauai, Hawaii, USA: IEEE, 2001, **2**: II-123−II-130

12 Ahad M A R, Tan J, Kim H, Ishikawa S. Action dataset-a survey. In: Proceedings of the 2011 SICE Annual Conference (SICE). Tokyo, Japan: IEEE, 2011. 1650−1655

13 Chaquet J M, Carmona E J, Fernández-Caballero A. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 2013, **117**(6): 633−659

14 Zhang J, Li W Q, Ogunbona P O, Wang P C, Tang C. RGB-D-based action recognition datasets: a survey. *Pattern Recognition*, 2016, **60**: 86−105

15 Aggarwal J K, Ryoo M S. Human activity analysis: a review. *ACM Computing Surveys*, 2011, **43**(3): Article No. 16

16 Vishwakarma S, Agrawal A. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 2013, **29**(10): 983−1009

17 Chen C, Jafari R, Kehtarnavaz N. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications*, 2017, **76**(3): 4405−4425

18 Shan Yan-Hu, Zhang Zhang, Huang Kai-Qi. Visual human action recognition: history, status and prospects. *Journal of Computer Research and Development*, 2016, **53**(1): 93−112
(单言虎, 张彰, 黄凯奇. 人的视觉行为识别研究回顾、现状及展望. 计算机研究与发展, 2016, **53**(1): 93−112)

19 Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR). Cambridge, UK: IEEE, 2004, **3**: 32−36

20 Blank M, Gorelick L, Shechtman E, Irani M, Basri R. Actions as space-time shapes. In: Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05). Beijing, China: IEEE, 2005, **2**: 1395−1402

21 Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(12): 2247−2253

22 Zhou T C, Li N J, Cheng X, Xu Q J, Zhou L, Wu Z Y. Learning semantic context feature-tree for action recognition via nearest neighbor fusion. *Neurocomputing*, 2016, **201**: 1−11

23 Xu W R, Miao Z J, Tian Y. A novel mid-level distinctive feature learning for action recognition via diffusion map. *Neurocomputing*, 2016, **218**: 185−196

24 Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes [Online], available: http://www.wisdom.weizmann.ac.il/∼vision/SpaceTimeActions.html, January 26, 2016.

25 Tran D, Sorokin A. Human activity recognition with metric learning. In: Proceedings of the 10th European Conference on Computer Vision (ECCV). Marseille, France: Springer, 2008. 548−561

26 Fathi A, Mori G. Action recognition by learning mid-level motion features. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, AK, USA: IEEE, 2008. 1−8

27 Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, AK, USA: IEEE, 2008. 1−8

28 Rodriguez M D, Ahmed J, Shah M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, AK, USA: IEEE, 2008. 1−8

29 Marszalek M, Laptev I, Schmid C. Actions in context. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, FL, USA: IEEE, 2009. 2929−2936

30 Liu J G, Luo J B, Shah M. Recognizing realistic actions from videos "in the wild". In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, FL, USA: IEEE, 2009. 1996−2003

31 Niebles J C, Chen C W, Li F F. Modeling temporal structure of decomposable motion segments for activity classification. In: Proceedings of the 11th European Conference on Computer Vision (ECCV): Part II. Heraklion, Crete, Greece: Springer, 2010. 392−405

32 Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: a large video database for human motion recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011. 2556−2563

33 Reddy K K, Shah M. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 2013, **24**(5): 971−981

34 Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012. 1−7

35 Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F F. Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 1725−1732

36 Kulkarni K, Evangelidis G, Cech J, Horaud R. Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 2015, **112**(1): 90−114

37 Shabani A H, Clausi D A, Zelek J S. Evaluation of local spatio-temporal salient feature detectors for human action recognition. In: Proceedings of the 2012 Ninth Conference on Computer and Robot Vision (CRV). Toronto, ON, Canada: IEEE, 2012. 468−475

38 Fernando B, Anderson P, Hutter M, Gould S. Discriminative hierarchical rank pooling for activity recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 1924−1932

39 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, Australia: IEEE, 2013. 3551−3558

40 Liu A A, Su Y T, Nie W Z, Kankanhalli M. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(1): 102−114

41 Wang Y, Tran V, Hoai M. Evolution-preserving dense trajectory descriptors. arXiv: 1702.04037, 2017.

42 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. *Advance in Neural Information Processing Systems*. 2014, **1**(4): 568−576

43 Al Harbi N, Gotoh Y. A unified spatio-temporal human body region tracking approach to action recognition. *Neurocomputing*, 2015, **161**: 56−64

44 Tong M, Wang H Y, Tian W J, Yang S L. Action recognition new framework with robust 3D-TCCHOGAC and 3D-HOOFGAC. *Multimedia Tools and Applications*, 2017, **76**(2): 3011−3030

45 Vishwakarma D K, Kapoor R, Dhiman A. Unified framework for human activity recognition: an approach using spatial edge distribution and $\Re$-transform. *AEU-International Journal of Electronics and Communications*, 2016, **70**(3): 341−353

46 Vishwakarma D K, Kapoor R, Dhiman A. A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. *Robotics and Autonomous Systems*, 2016, **77**: 25−38

47 Liu C W, Pei M T, Wu X X, Kong Y, Jia Y D. Learning a discriminative mid-level feature for action recognition. *Science China Information Sciences*, 2014, **57**(5): 1−13

48 Laptev I, Marszalek M, Schmid C, Rozenfeld B. Hollywood2: Human actions and scenes dataset [Online], available: http://www.di.ens.fr/ laptev/actions/hollywood2/, March 12, 2016.

49 Wang H, Kläser A, Schmid C, Liu C L. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 2013, **103**(1): 60−79

50 Soomro K, Zamir A R. Action recognition in realistic sports videos. *Computer vision in sports*. Cham, Switzerland: Springer, 2014. 181−208

51 Peng X J, Zou C Q, Qiao Y, Peng Q. Action recognition with stacked fisher vectors. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 581−595

52 Liu C H, Liu J, He Z C, Zhai Y J, Hu Q H, Huang Y L. Convolutional neural random fields for action recognition. *Pattern Recognition*, 2016, **59**: 213−224

53 Sun Q R, Liu H, Ma L Q, Zhang T W. A novel hierarchical bag-of-words model for compact action representation. *Neurocomputing*, 2016, **174**(Part B): 722−732

54 Sekma M, Mejdoub M, Amar C B. Human action recognition based on multi-layer fisher vector encoding method. *Pattern Recognition Letters*, 2015, **65**(C): 37−43

55 Li Y W, Li W X, Mahadevan V, Vasconcelos N. VLAD^3: encoding dynamics of deep features for action recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 1951−1960

56 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 1933−1941

57 Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O, Van Gool L. Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, the Netherlands: Springer, 2016. 20−36

58 Wang H S, Wang W, Wang L. How scenes imply actions in realistic videos? In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016. 1619−1623

59 Wang L M, Guo S, Huang W L, Qiao Y. Places205-VGGNet models for scene recognition. arXiv: 1508.01667, 2015.

60 Lan Z Z, Lin M, Li X C, Hauptmann A G, Raj B. Beyond Gaussian pyramid: multi-skip feature stacking for action recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 204−212

61 Ijjina E P, Chalavadi K M. Human action recognition using genetic algorithms and convolutional neural networks. *Pattern Recognition*, 2016, **59**: 199−212

62 Lev G, Sadeh G, Klein B, Wolf L. RNN Fisher vectors for action recognition and image annotation. In: Proceedings of the 14th European Conference on Computer Vision (ECCV): Part VIII. Amsterdam, the Netherlands: Springer, 2016. 833−850

63 Jiang Y G, Liu J G, Zamir A R, Laptev I, Piccardi M, Shah M, Sukthankar R. THUMOS challenge: Action recognition with a large number of classes [Online], available: http://crcv.ucf.edu/ICCV13-Action-Workshop/index.html, November 20, 2016.

64 Jiang Y G, Liu J G, Zamir A R, Toderici G, Laptev I, Shah M, Sukthankar R. THUMOS challenge: action recognition with a large number of classes [Online], available: http://crcv.ucf.edu/THUMOS14/home.html, November 20, 2016.

65 Gorban A, Idrees H, Jiang Y G, Zamir A R, Laptev I, Shah M, Sukthankar R. THUMOS challenge: action recognition with a large number of classes [Online], available: http://www.thumos.info/home.html, November 20, 2016.

66 Xu Z, Zhu L, Yang Y, Hauptmann A G. UTS-CMU at THUMOS 2015. In: Proceedings of the 2015 THUMOS Challenge. Boston, MA, USA: CVPR, 2015. 1−3

67 Mahasseni B, Todorovic S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 3054−3062

68 Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 2006, **104**(2−3): 249−257

69 Singh S, Velastin S A, Ragheb H. MuHAVi: a multicamera human action video dataset for the evaluation of action recognition methods. In: Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Boston, MA, USA: IEEE, 2010. 48−55

70 Ferryman J, Shahrokni A. PETS2009: dataset and challenge. In: Proceedings of the 22th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter). Snowbird, UT, USA: IEEE, 2009. 1−6

71 Patino L, Ferryman J. PETS 2014: dataset and challenge. In: Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Seoul, South Korea: IEEE, 2014. 355−360

72 Ashraf N, Foroosh H. Motion retrieval using consistency of epipolar geometry. In: Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP). Quebec City, QC, Canada: IEEE, 2015. 4219−4223

73 Ji X F, Ju Z J, Wang C, Wang C H. Multi-view transition HMMs based view-invariant human action recognition method. *Multimedia Tools and Applications*, 2016, **75**(19): 11847−11864

74 Gao Z, Nie W Z, Liu A N, Zhang H. Evaluation of local spatial-temporal features for cross-view action recognition. *Neurocomputing*, 2016, **173**(Part 1): 110−117

75 Wu D, Shao L. Multi-max-margin support vector machine for multi-source human action recognition. *Neurocomputing*, 2014, **127**(3): 98−103

76 Yi Y, Lin M Q. Human action recognition with graph-based multiple-instance learning. *Pattern Recognition*, 2016, **53**(C): 148−162

77 Jung H J, Hong K S. Modeling temporal structure of complex actions using bag-of-sequencelets. *Pattern Recognition Letters*, 2017, **85**: 21−28

78 Ballas N, Yang Y, Lan Z Z, Delezoide B, Preteux F, Hauptmann A. Space-time robust representation for action recognition. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, NSW, Australia: IEEE, 2013. 2704−2711

79 Qiu Z F, Li Q, Yao T, Mei T, Rui Y. MSR Asia MSM at THUMOS challenge 2015. In: Proceedings of the 2015 THUMOS Challenge. Boston, MA, USA: CVPR, 2015. 1−3

80 Ning K, Wu F. ZJUDCD submission at THUMOS challenge 2015. In: Proceedings of the 2015 THUMOS Challenge. Boston, MA, USA: CVPR, 2015. 1−2

81 Ng J Y H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: deep networks for video classification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 4694−4702

82 Moghaddam Z, Piccardi M. Training initialization of Hidden Markov Models in human action recognition. *IEEE Transactions on Automation Science and Engineering*, 2014, **11**(2): 394−408

83 Wu X X, Jia Y D. View-invariant action recognition using latent kernelized structural SVM. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 411−424

84 Alcantara M F, Moreira T P, Pedrini H. Real-time action recognition using a multilayer descriptor with variable size. *Journal of Electronic Imaging*, 2016, **25**(1): Article No. 013020

85 Chaaraoui A A, Flórez-Revuelta F. Human action recognition optimization based on evolutionary feature subset selection. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation. Amsterdam, the Netherlands: ACM, 2013. 1229−1236

86 Cai J X, Tang X, Feng G C. Learning pose dictionary for human action recognition. In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR). Stockholm, Sweden: IEEE, 2014. 381−386

87 Chaaraoui A A, Flórez-Revuelta F. A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views. *International Scholarly Research Notices*, 2014, **2014**: Article No. 547069

88 Alcantara M F, Moreira T P, Pedrini H. Real-time action recognition based on cumulative motion shapes. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014. 2917−2921

89 Li L Z, Nawaz T, Ferryman J. PETS 2015: datasets and challenge. In: Proceedings of the 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Karlsruhe, Germany: IEEE, 2015. 1−6

90 Patino L, Cane T, Vallee A, Ferryman J. PETS 2016: dataset and challenge. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Las Vegas, NV, USA: IEEE, 2016. 1240−1247

91 PETS 2014 [Online], available: http://www.cvg.reading.ac.uk/PETS2014/, April 16, 2016

92 Chen J W, Wu J, Konrad J, Ishwar P. Semi-coupled two-stream fusion ConvNets for action recognition at extremely low resolutions. In: Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, California, USA: IEEE, 2017. 139−147

93 Yang A Y, Jafari R, Sastry S S, Bajcsy R. Distributed recognition of human actions using wearable motion sensor networks. *Journal of Ambient Intelligence and Smart Environments*, 2009, **1**(2): 103−115

94 CMU graphics lab motion capture database [Online], available: http://mocap.cs.cmu.edu, September 27, 2016.

95 Li W Q, Zhang Z Y, Liu Z C. Action recognition based on a bag of 3D points. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). San Francisco, CA, USA: IEEE, 2010. 9−14

96 Wang J, Liu Z C, Wu Y, Yuan J S. Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012. 1290−1297

97 Ellis C, Masood S Z, Tappen M F, LaViola Jr J J, Sukthankar R. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 2013, **101**(3): 420−436

98 Yang A Y, Iyengar S, Kuryloski P, Jafari R. Distributed segmentation and classification of human actions using a wearable motion sensor network. In: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08). Anchorage, AK, USA: IEEE, 2008. 1−8

99 Guo Y C, He W H, Gao C. Human activity recognition by fusing multiple sensor nodes in the wearable sensor systems. *Journal of Mechanics in Medicine and Biology*, 2012, **12**(5): Article No. 1250084

100 Guo M, Wang Z L. A feature extraction method for human action recognition using body-worn inertial sensors. In: Proceedings of the 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Calabria, Italy: IEEE, 2015. 576−581

101 Jia Q, Fan X, Luo Z X, Li H J, Huyan K, Li Z Z. Cross-view action matching using a novel projective invariant on non-coplanar space-time points. *Multimedia Tools and Applications*, 2016, **75**(19): 11661−11682

102 Al Aghbari Z, Junejo I N. DisCoSet: discovery of contrast sets to reduce dimensionality and improve classification. *International Journal of Computational Intelligence Systems*, 2015, **8**(6): 1178−1191

103 Kadu H, Kuo C C J. Automatic human Mocap data classification. *IEEE Transactions on Multimedia*, 2014, **16**(8): 2191−2202

104 Wang J, Nie X H, Xia Y, Wu Y, Zhu S C. Cross-view action modeling, learning, and recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 2649−2656

105 Chen C, Jafari R, Kehtarnavaz N. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP). Quebec City, QC, Canada: IEEE, 2015. 168−172

106 Shahroudy A, Liu J, Ng T T, Wang G. NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 1010−1019

107 Luo J J, Wang W, Qi H R. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV). Sydney, NSW, Australia: IEEE, 2013. 1809−1816

108 Chen C, Jafari R, Kehtarnavaz N. Action recognition from depth sequences using depth motion maps-based local binary patterns. In: Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, 2015. 1092−1099

109 Chen W B, Guo G D. Triviews: a general framework to use 3D depth data effectively for action recognition. *Journal of Visual Communication and Image Representation*, 2015, **26**: 182−191

110 Wang P C, Li W Q, Gao Z M, Zhang J, Tang C, Ogunbona P. Deep convolutional neural networks for action recognition using depth map sequences. arXiv: 1501.04686, 2015. 1−8

111 Zhang H L, Zhong P, He J L, Xia C X. Combining depth-skeleton feature with sparse coding for action recognition. *Neurocomputing*, 2017, **230**: 417−426

112 Shahroudy A, Ng T T, Gong Y H, Wang G. Deep multimodal feature analysis for action recognition in RGB+D videos. arXiv: 160307120, 2016.

113 Kerola T, Inoue N, Shinoda K. Cross-view human action recognition from depth maps using spectral graph sequences. *Computer Vision and Image Understanding*, 2017, **154**: 108−126

114 Beh J, Han D K, Durasiwami R, Ko H. Hidden Markov model on a unit hypersphere space for gesture trajectory recognition. *Pattern Recognition Letters*, 2014, **36**: 144−153

115 Liu M Y, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 2017, **68**: 346−362

116 Li C K, Hou Y H, Wang P C, Li W Q. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 2017, **24**(5): 624−628

117 Bulbul M F, Jiang Y S, Ma J W. DMMs-based multiple features fusion for human action recognition. *International Journal of Multimedia Data Engineering & Management*, 2015, **6**(4): 23−39

118 Wang P C, Li W Q, Li C K, Hou Y H. Action recognition based on joint trajectory maps with convolutional neural networks. arXiv: 1612.09401v1, 2016. 1−11

119 Kwolek B, Kepski M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 2014, **117**(3): 489−501

120 Gasparrini S, Cippitelli E, Spinsante S, Gambi E. A depth-based fall detection system using a kinect? sensor. *Sensors*, 2014, **14**(2): 2756−2775

121 Gasparrini S, Cippitelli E, Gambi E, Spinsante S, Wåhslén J, Orhan I, Lindh T. Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. *ICT innovations 2015*. Cham, Switzerland: Springer, 2016. 99−108

122 Su Ben-Yue, Jiang Jing, Tang Qing-Feng, Sheng Min. Human dynamic action recognition based on functional data analysis. *Acta Automatica Sinica*, 2017, **43**(5): 866−876 (苏本跃, 蒋京, 汤庆丰, 盛敏. 基于函数型数据分析方法的人体动态行为识别. 自动化学报, 2017, **43**(5): 866−876)

123 Han L, Wu X X, Liang W, Hou G M, Jia Y D. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 2010, **28**(5): 836−849

124 Wang J, Liu Z C, Wu Y, Yuan J S. Learning actionlet ensemble for 3D human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(5): 914−927

125 Chen H Z, Wang G J, Xue J H, He L. A novel hierarchical framework for human action recognition. *Pattern Recognition*, 2016, **55**: 148−159

126 Zhu Y, Chen W B, Guo G D. Fusing multiple features for depth-based action recognition. *ACM Transactions on Intelligent Systems and Technology*, 2015, **6**(2): Article No. 18

127 Jiang X B, Zhong F, Peng Q S, Qin X Y. Robust action recognition based on a hierarchical model. In: Proceedings of the 2013 International Conference on Cyberworlds (CW). Yokohama, Japan: IEEE, 2013. 191−198

128 Chen C C, Aggarwal J K. Recognizing human action from a far field of view. In: Proceedings of the 2009 Workshop on Motion and Video Computing (WMVC′09). Snowbird, UT, USA: IEEE, 2009. 1−7

129 Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints. In: Proceedings of the 12th International Conference on Computer Vision (ICCV). Kyoto, Japan: IEEE, 2009. 104−111

130 Ryoo M S, Aggarwal J K. UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA) [Online], available: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, December 10, 2016.

131 Jiang Y G, Ye G N, Chang S F, Ellis D, Loui A C. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR′11). Trento, Italy: ACM, 2011. Article No. 29

132 Rohrbach M, Amin S, Andriluka M, Schiele B. A database for fine grained activity detection of cooking activities. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012. 1194−1201

133 Rohrbach M, Regneri M, Andriluka M, Amin S, Pinkal M, Schiele B. Script data for attribute-based recognition of composite activities. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 144−157

134 Bojanowski P, Lajugie R, Bach F, Laptev I, Ponce J, Schmid C, Sivic J. Weakly supervised action labeling in videos under ordering constraints. *Computer Vision — ECCV 2014*. Cham, Germany: IEEE, 2014, **8693**: 628−643

135 Rohrbach M, Rohrbach A, Regneri M, Amin S, Andriluka M, Pinkal M, Schiele B. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 2016, **119**(3): 346−373

136 Heilbron F C, Escorcia V, Ghanem B, Niebles J C. Activitynet: a large-scale video benchmark for human activity understanding. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 961−970

137 Gkalelis N, Kim H, Hilton A, Nikolaidis N, Pitas I. The i3DPost multi-view and 3D human action/interaction database. In: Proceedings of the 2009 Conference for Visual Media Production (CVMP). London, UK: IEEE, 2009. 159−168

138 De la Torre F, Hodgins J K, Montano J, Valcarcel S. Detailed human data acquisition of kitchen activities: the CMU-multimodal activity database (CMU-MMAC). In: Proceedings of the 2009 Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research, in Conjuction with CHI. Boston, MA, USA: ACM, 2009. 1−5

139 Ni B B, Wang G, Moulin P. RGBD-HuDaAct: a color-depth video database for human daily activity recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona, Spain: IEEE, 2011. 1147−1153

140 Xia L, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3D joints. In: Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Providence, RI, USA: IEEE, 2012. 20−27

141 Cheng Z W, Qin L, Ye Y T, Huang Q Q, Tian Q. Human daily action analysis with multi-view and color-depth data. In: Proceedings of the Computer Vision, ECCV 2012-Workshops and Demonstrations. Florence, Italy: Springer, 2012. 52−61

142 Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R. Berkeley MHAD: a comprehensive multimodal human action database. In: Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV). Tampa, FL, USA: IEEE, 2013. 53−60

143 Oreifej O, Liu Z C. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013. 716−723

144 Wei P, Zhao Y B, Zheng N N, Zhu S C. Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1165−1179

145 Yu G, Liu Z C, Yuan J S. Discriminative orderlet mining for real-time recognition of human-object interaction. In: Proceedings of the 12th Asian Conference on Computer Vision (ACCV). Singapore: Springer, 2014. 50−65

146 Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S. YouTube-8M: a large-scale video classification benchmark. arXiv: 1609.08675, 2016. 1−10

朱红蕾 兰州理工大学计算机与通信学院博士研究生. 2004 年获得兰州理工大学硕士学位. 主要研究方向为计算机视觉与模式识别. 本文通信作者.
E-mail: zhuhllut@139.com
(**ZHU Hong-Lei** Ph. D. candidate at the School of Computer and Conmunacation, Lanzhou University of Technology. She received her master degree from Lanzhou University of Technology in 2004. Her research interest covers computer vision and pattern recognition. Corresponding author of this paper.)

朱昶胜 兰州理工大学计算机与通信学院教授. 2006 年获得兰州理工大学博士学位. 主要研究方向为高性能计算, 数据分析与理解.
E-mail: zhucs_2008@163.com
(**ZHU Chang-Sheng** Professor at the School of Computer and Conmunacation, Lanzhou University of Technology. He received his Ph. D. degree from Lanzhou University of Technology in 2006. His research interest covers high performance computing, data analysis, and understanding.)

徐志刚 兰州理工大学计算机与通信学院副教授. 2012 年获得中国科学院研究生院博士学位. 主要研究方向为计算机视觉与图像处理.
E-mail: xzg_cn@163.com
(**XU Zhi-Gang** Associate professor at the School of Computer and Conmunacation, Lanzhou University of Technology. He received his Ph. D. degree from Graduate University of Chinese Academy of Sciences in 2012. His research interest covers computer vision and image processing.)