第44卷 第10期 北 京 工 业 大 学 学 报 Vol. 44 No. 10
2018年 10月 JOURNAL OF BEIJING UNIVERSITY OF TECHNOLOGY Oct. 2018

基于深度图像的人体行为识别综述

孙 彬，孔德慧，张雯晖，贾文浩
（北京工业大学信息学部多媒体与智能软件技术北京市重点实验室，北京 100124）

摘 要：深度图像降低了人体三维运动信息在视觉获取过程中的维度损失，使得与传统彩色图像相比，基于深度图像的人体行为识别研究在特征提取、表示及识别精度等方面体现出技术优势，受到广泛关注，因此，全面、深入地综述了基于深度图像的人体行为识别的研究现状. 首先，对近年来提出的基于深度图像的人体行为识别的各种方法进行整理、分类；然后，对多个常用的人体行为公开数据库进行介绍，并在3个数据库上对不同方法的识别率进行对比分析；最后，阐述了人体行为识别技术未来可能的发展趋势.

关键词：人体行为识别；特征提取；深度图像；人体关节点；机器学习
中图分类号：TP 391 **文献标志码**：A **文章编号**：0254－0037(2018)10－1353－16
doi：10.11936/bjutxb2017040051

Survey on Human Action Recognition From Depth Maps

SUN Bin, KONG Dehui, ZHANG Wenhui, JIA Wenhao
(Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Depth maps reduce the dimension loss of 3D human motion information in the process of vision acquisition, therefore depth map-based human action recognition reflects technical advantages in fields of feature extraction, representation and recognition accuracy, compared with traditional RGB image, and attracts the extensive attention. The research status of depth map-based human action recognition was summarized in this paper. First, the existing methods of recognizing human action from depth maps were collated and classified. Then, multiple publicly available human action datasets were introduced, and the accuracies of several datasets in different methods were compared. Finally, the possible future directions of human action recognition were analyzed.

Key words: human action recognition; feature extraction; depth map; body joint; machine learning

基于计算机视觉技术的人体行为识别在人类生活的众多领域得到广泛应用，如视频监控、运动检索、人机交互、智能家居以及医疗保健等[1-4]. 传统的基于彩色(RGB)相机获取视频序列进行行为识别的方法[5-10]有很多，如基于时空特征的方法[11-12]和基于运动轨迹特征的方法[1,13-14]. 但是，基于RGB信息的人体行为识别具有多方面的挑战：1) 复杂背景、遮挡、阴影、尺度变化以及不同的光照条件都会对识别带来很大的困难，这也是基于RGB行为识别的难点；2) 同样的动作从不同的视角会生成不同的视图；3) 不同的人表演同一个动作会有很大的不同，并且2个不同的动作类又可能会有很大的相似性. RGB视觉信息所存在的这些固有缺陷限制了基于RGB信息的人体行为识别的性能.

收稿日期：2017-04-27
基金项目：国家自然科学基金资助项目(61370120)；北京市自然科学基金资助项目(4162009)
作者简介：孙 彬(1990—)，男，博士研究生，主要从事模式识别方面的研究，E-mail: sunbin1357@emails.bjut.edu.cn
通信作者：孔德慧(1968—)，女，教授，博士生导师，主要从事模式识别、虚拟现实与图形学方面的研究，E-mail: kdh@bjut.edu.cn

近几年,随着传感器技术迅速发展,高清的深度相机逐渐普及,例如 Microsoft Kinect. Kinect 的成本低,尺寸小,并且可以很容易地获得高分辨率的深度图像(depth map) . 深度图像中的每个像素记录的是场景的深度值,而不是光强度. 深度相机的引入拓展了计算机系统感知 3D 视觉世界的能力,在一定程度上弥补了感知过程中将三维对象信息捕获为二维视觉信息时的维度信息缺失. 与 RGB 视觉信息相比,深度图像通过所提供的场景的结构信息可以极大地减轻遮挡、复杂背景等因素的影响,并且在不同的光照条件下,颜色和纹理具有不变性. 从单个视角,如果不同的行为有相似的 2D 投影,深度图像可以提供额外的体形信息来区分不同的行为. 此外,Kinect 还提供了强大的骨架追踪算法[15],该方法可以实时输出每一帧 3D 人体关节点的位置. 人体的骨架关节点不会受尺度、视角变化的影响,因此,利用 Kinect 所提供的人体骨架关节点进行行为识别是一个有前途的研究方向.

近几年来,在基于深度图像的人体行为识别方面,研究者们以提取更具有行为区分能力的人体运动深度特征作为核心问题,提出了大量的基于深度图像的人体行为识别的方法,并构建了多种人体行为深度图像数据库以评估识别效果. 本文将综述基于深度图像的识别方法,并介绍相关深度图像数据库,最后对人体行为识别技术的发展趋势进行探讨.

1 特征提取及表示方法

解决识别问题的一般流程包括 2 个环节: 基于输入信息的特征提取以及基于特征表示的对象分类. 因此,采用这一模式的人体行为识别方法,无论是基于彩色图像,还是基于深度图像,仅就对象分类的层面而言,二者之间不存在本质差异. 而由于输入数据形式的不同所导致的特征提取与表示方面的差异,是使 2 种识别方法产生本质差异的关键因素. 因此,本文将对基于深度图像的人体行为识别方法在特征提取与表示方法方面集中展开论述.

深度图像作为一种包含深度信息的三维空间平面投影图,其优势在于以 2.5D 的形式提供了对象的空间几何信息,基于深度图像的人体行为特征提取与表示也势必围绕几何要素来实现. 对现有识别方法所提取特征进行分析,根据其所对应的信息维度可大体分为 4 类: 点特征、线特征、面特征和体特征(如表 1 所示) .

基于点特征的方法是针对点(关节点、兴趣点等) 提取特征; 基于线特征的方法是根据点和点之间的关系所构成的线来提取特征; 基于面特征的方法主要是通过在曲面上计算曲面法线来提取特征; 基于体特征的方法主要是通过在(x,y,z) 的三维体或(x,y,z,t) 四维时空体提取特征. 上述 4 类特征的共同点在于均在某种意义下体现了人体行为的几何不变性,适用于进行人体特征表示与分类; 而其差异的本质在于信息关联程度的变化,即信息自由度的不同.

表 1 4 种特征的含义

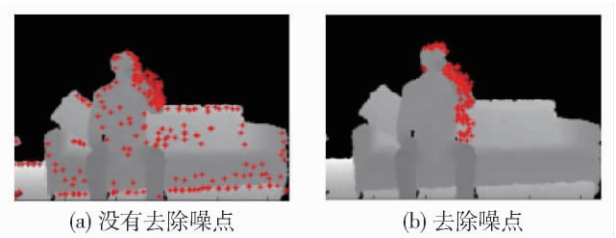
Table 1 Meanings of the four features

名称	含义
点特征	表示深度图像中的关节点、兴趣点信息
线特征	在点特征的基础上建立点点关联信息
面特征	利用点特征估算相应表面的法线信息
体特征	在深度图像(序列) 的 3D(4D) 重建空间进行特征提取

1.1 基于点特征的的方法

对 RGB 信息提取时空兴趣点(spatio temporal interest points,STIPs) 特征已经被证明是一种有效的描述方法,它将人体动作信息以一些不关联的点的形式进行描述. 兴趣点通过描述场景的局部,提供了图像内容的紧凑表示,这样增强了对混乱、遮挡和类内差异的鲁棒性. 目前,存在很多检测时空兴趣点和计算局部特征描述子的方法[16]. 使用比较多的兴趣点的检测方法包括 Harris3D 检测[17]、Hessian 检测[18]和 cuboid 检测[19]. 提取特征描述子的方法包括方向梯度直方图(histogram of oriented gradient, HOG) [20]、尺度不变特征变换(scale invariant feature transform, SIFT) [21]、加速鲁棒特征(speed up robust feature, SURF) [18]和核描述子[22]. 近几年,国内外研究学者将提取时空兴趣点的方法应用到深度图像中. Zhu 等[23]尝试了不同兴趣点检测和特征描述子的结合. Ni 等[24]使用了 Harris3D 检测和 HOG/光流场方向直方图(histogram of oriented optical flow,HOF) 描述子进行行为识别. 文献[25-27]也采用了 Harris3D 检测提取时空兴趣点,不同的是 Zhao 等[25]结合了 HOG/HOF 和局部深度模式(local depth pattern,LDP) 表示特征. LDP 特征用兴趣点作为局部块(patch) 的中心. 局部块的大小与兴趣点的深度值有关. 每一个局部块都会被分成网格(grid) ,再对每一个网格计算平均深度

值，并计算 2 个网格平均深度值的差. 这种差的特征向量就是所提出的 LDP 特征. Chen 等[26]则是将深度图像映射到 3 个正交面，在每个面提取时空兴趣点、轨迹形状和运动边界. 并使用了 HOG、HOF 和运动边界直方图(motion boundary histograms，MBH) 描述子. Cheng 等[27]通过计算中心点与附近的 26 个点的深度值来构造比较编码描述子(comparative coding descriptor，CCD) . Harris3D 检测等方法本身是针对 RGB 图形的，但是深度图像中又有很多的噪声. 为了解决这个问题，Xia 等[28]使用滤波方法从深度视频提取 STIPs (depth STIPs，DSTIPs) ，并使用深度立方体相似特征(depth cuboid similarity feature，DCSF) 描述局部 3D 深度立方体. DSTIPs 可以有效地抑制噪声(如图 1 所示) ，DCSF 则是基于自相似性来描述 3D 立方体的时空形状.



(a) 没有去除噪点　　(b) 去除噪点

图 1　DSTIPs[28] x-y 面上的结果

Fig. 1　DSTIPs[28] projection in the x-y plane

Xia 等[29]提出了一种基于 3D 关节点位置直方图的表示方法，通过修正的球坐标系统将 3D 空间分成 84 个空间(bins) ，然后将关节点位置投影到这 84 个 bins 上构成直方图，并使用线性判别分析(linear discriminant analysis，LDA) 对特征进行降维并聚类成 k 个姿态视觉单词，再将深度序列编码到连续的单词中，最后通过隐马尔可夫模型(hidden Markov model，HMM) 分类器进行分类. Salih 等[30]首先提取人体关节点的球面角，再将每一帧关节点的球面角投影到修正球谐(modified spherical harmonics，MSHs) 的基函数，并用 MSHs 的协方差作为视频序列的描述子.

1.2　基于线特征的的方法

Yang 等[31]提出了一种基于特征关节点(EigenJoints) 的行为识别方法(如图 2 所示) . 该方法通过计算 3D 关节点之间的位置关系来描述一个动作序列，包括静态的姿态特征、连续的运动特征和偏移特征. 静态的姿态特征表示当前帧关节点间的位置差，连续的运动特征表示关节点在当前帧与前一帧的位置差，而偏移特征则是通过计算关节点在当前帧与初始帧的位置差构成，这三通道特征的组合构成了初步的特征表示. 然后，对这些特征进行归一化，通过主成分分析法(principal components analysis，PCA) 降维来减少冗余度和噪声. 最后，使用朴素贝叶斯最近邻(naive bayes nearest neighbor，NBNN) 分类器进行分类.

文献[32-33]也都使用了静态的姿态特征，不同的是 Li 等[32]选择了关节点间距离的相对变化最大的 K 个关节点对，并将其构成关节点空间图. 另外，使用时域金字塔协方差描述子表示关节点空间图中关节点的位置，并使用训练的图核(graph kernel) 和支持向量机(support vector machine，SVM) 进行分类. Luo 等[33]使用稀疏编码的方法学习静态特征，并用基于最大池化(max pooling) 的时域金字塔结构对特征进行直方图表示，最后采用 SVM 进行分类. 文献[34]计算了连续的运动特征和静态的姿态特征，不过 Jiang 等[34]所提取的静态的姿态特征首先选择一个关节点作为基点，然后计算每个关节点与基点的位置差. 该方法还通过一种加权图描述两通道的特征，这种加权图可以处理关节点不稳定、序列

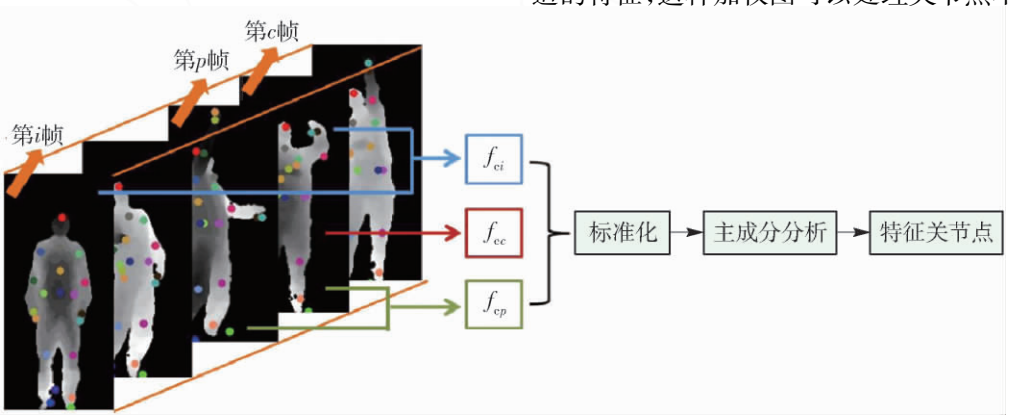


图 2　关节点特征提取与表示的过程[31]

Fig. 2　Process of feature extraction and representation of joints[31]

长度不同等问题. 最后使用序列匹配的方法进行分类.

Zanfir 等[35]仅将关节点的位置作为静态特征,然后以当前帧为中心使用 5 帧的时间窗口计算速度和加速度特征,最后采用改进的 K 最近邻(K-nearest neighbor,KNN) 进行分类. 文献[36]也采用了速度、加速度 2 个特征,不同的是 Cai 等[36]将每一帧计算四肢的方向作为静态特征,然后通过马尔科夫随机场编译这些肢体的特征来减少类内的差距,最后使用多通道多实例学习算法学习有判别力的骨架运动. Lu 等[37]没有采用静态特征,而是提出了另外一种关节点位置关系来描述一个动作序列,该方法通过计算一段时间范围内 3D 关节点的位置偏移作为动作序列的特征描述,然后采用词袋(bag of word,BOW) 的方法来量化特征,最后采用 NBNN 进行分类.

1.3 基于面特征的方法

Tang 等[38]通过法线向量方向直方图(histogram of oriented normal vectors,HONV) 估计了 3D 形状曲面上每个点的 3D 法线向量,并在每一个局部块(patch) 构建了 2D 法线向量直方图. 该方法证明了曲面法线可以提供很多 3D 物体的形状和结构信息. 受该方法的启发,Oreifej 等[39]采用了 4D 曲面法线的方向直方图(histogram of oriented 4D normals,HON4D) . 与 3D 法线向量相比,4D 法线向量可以捕获形状和运动信息. 该方法将深度序列看作一个 4D 时空体,计算每个点云的 4D 法线,并构建了一个 4D 法线向量的直方图(如图 3 所示) . 为了构建 HON4D,在 4D 空间使用 600 个单元、120 个顶点的四维体进行初始量化,并将该 4D 空间的曲面法线投影到顶点上. HON4D 特征通过训练数据学习了具有判别性的非均匀投影子(projector) ,从而使用了非均匀的 4D 量化. 非均匀的 4D 量化与均匀的 4D 量化相比表现出更好的性能. 为了保持相邻法线的关系,使它们对噪声有更好的鲁棒性,Yang 等[40]聚集超曲面法线来描述局部运动和形状信息. 为了能够获得全局的时空特性,该方法还提出了用自适应时空金字塔将深度视频细分成一组时空网格. 每一个网格的特征向量联合起来则是最后的超曲面法向量(super normal vector, SNV) 表示模型. 不同于上述在点云上求曲面法线,Zhou 等[41]则提出在骨架上求曲面法线. 在每一帧,整个人体根据骨架分成几个局部近似刚性的部分(local approximate rigid parts,LARPs) . 在每一个刚性部分用曲面法线来表示 LARPs,相邻帧对应的 LARPs 形成了全局近似刚性部分(global approximate rigid part, GARP) . 此外,骨架时序的变化用来捕获 GARP 时序的几何结构.

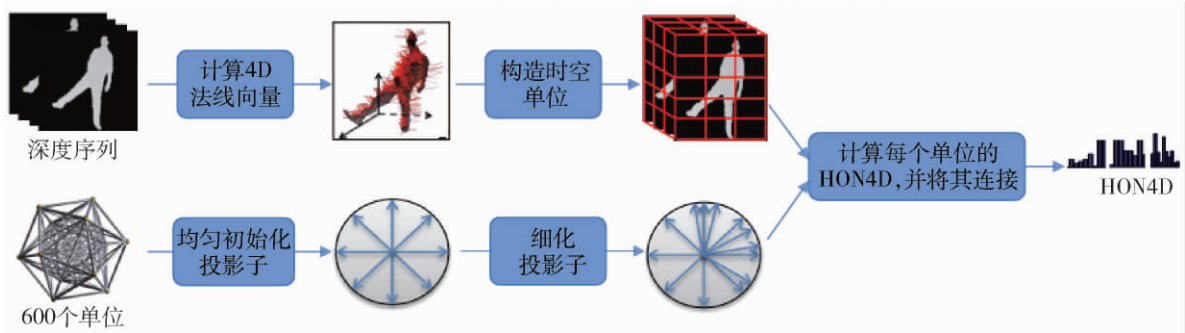


图 3 计算 HON4D 的步骤[39]

Fig. 3 Steps for computing HON4D[39]

1.4 基于体特征的方法

虽然在深度序列上应用 2D 视频特征可以获得较好的性能,但深度图像本身是一个 3D 的形体; 所以可以用 3D 的特性更好地描述人体动作,这样带有时间的深度视频序列就形成了一个 4D 时空体. Wang 等[42]基于每个关节点周围 3D 点云的局部占有信息提出了局部占有模式(local occupancy pattern,LOP) . 该方法将每一个关节点的局部空间划分成 $N_xN_yN_z$ 的空间网格,然后计算落入网格每个单元(bin) 中的点云数量,并采用 Sigmoid 函数获得每个单元的占有特征. 这种方法是在(x,y,z) 空间计算 LOP,Wang 等[43]又在(x,y,z,t) 这一 4D 时空体提出了随机占有模式(random occupancy pattern, ROP) 进行行为识别. 该方法通过在不同尺寸和不同位置随机采样四维体得到不同的子体(subvolume) ,并计算每个子体中点云的个数. 然后

将这些提取的特征进行稀疏编码，随后通过 SVM 进行识别. Vieira 等[44]并不是随机采样，而是提出了时空占有模式(space-time occupancy pattern，STOP)描述人体行为的 4D 时空体. 该方法将时空坐标均匀地分成多个部分来定义每个深度序列的 4D 网格. 假设人是静止的，通过计算每一个网格的占有信息可以粗略地描述人体姿态和人物交互. 也就是说，将每一个时间段的所有帧都放入同一空间中，计算每个网格中点云的个数(如图 4 所示). 另外，由于人体移动部分的网格比较稀疏，作用又不可忽视，所以该方法还使用了饱和的方案.

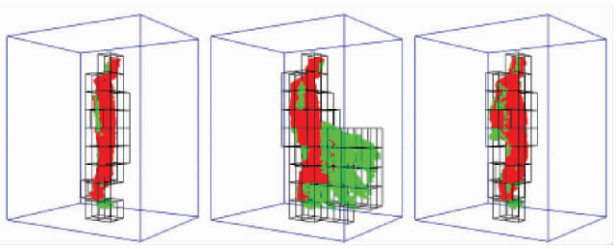


图 4　前踢动作深度序列的时空网格(红色点是饱和网格中的点)[44]

Fig. 4　Space-time cells of a depth sequence for the action forward kick (the red points for the saturated cells)[44]

1.5　基于融合的方法

每一种特征提取的方法都有自己的优点，并且相互独立. 如果能够将不同的特征进行有效地融合，进而获得一个更具判别性的特征向量，必然会提高识别性能. 因此，近年来融合的方法得到了学者的关注，并且已经在许多领域展示出很好的性能，比如多传感器系统[45]、多媒体分析[46]和人脸识别[47]等. 融合的方法总的来说分 2 种情况：特征层融合和决策层融合.

1.5.1　基于特征层融合的方法

特征层融合是早期的融合方法. 该方法首先通过不同的方法提取特征向量，然后将这些提取的特征进行标准化、选择或者转换，从而组合生成一个更具判别性的新的特征向量(如图 5 所示). Ohn-Bar 等[48]通过对深度图像提取改进的 HOG 特征，并提取关节点的角度特征进行融合，然后使用 SVM 进行分类. Zhu 等[23]则是对深度图像提取 STIP，并从关节点提取 EigenJoints[31]，然后使用随机森林进行融合和分类. 为了融合更多的特征，Zhu 等[49]又对 STIPs、时空自相关梯度(space-time auto-correlation of gradients，STACOG)[50]、EigenJoints[31]和方向 4D 曲面法线直方图(histogram of oriented 4D surface normals，HON4D)[39]4 个特征进行分析，做了各种融合方法的系统研究，使用了随机森林、联合互信息(joint mutual information)、条件互信息最大化(conditional mutual info maximization)等特征层融合的方法. 与上述方法不同，Luo 等[33]从关节点提取了每一帧的相对位置特征，又从 RGB 序列提取中心对称运动局部三元模式(center-symmetric motion local ternary pattern，CS-Mltp)特征. 将 2 种特征生成的直方图简单串联起来就形成了一个长的直方图，可以对其进行下一步分类.

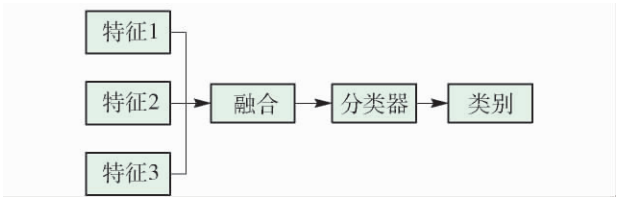


图 5　特征层融合的方法

Fig. 5　Fusion method of feature-level

1.5.2　基于决策层融合的方法

决策层融合与特征层融合方法不同，决策层融合首先通过每一种方法训练得到的分类器输出分类结果，然后将获得的分类结果进行融合，从而得出最后的分类结果(如图 6 所示). 文献[33，49，51-52]都是对每个特征所使用的分类器求出先验概率，不同的是文献[51-52]给每个特征所得到概率一个权重之后进行求和. Luo 等[33]采用了 2 种融合方法，使用求和运算将不同特征的概率直接求和，又使用乘积运算将不同特征的概率相乘. 而 Zhu 等[49]使用了投票、朴素贝叶斯结合、基于规则的融合(包括求和、最小值、最大值、中间值等)、多智能体系统(multiagent system)和基于 SVM 的融合等多种决策层融合的方法.

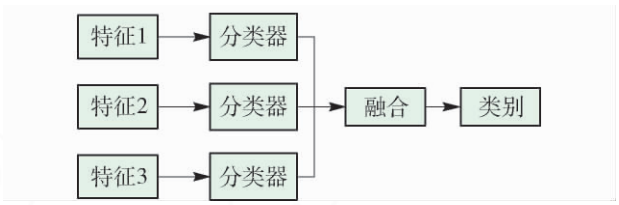


图 6　决策层融合的方法

Fig. 6　Fusion method of decision-level

1.6　基于机器学习的方法

机器学习的方法在计算机视觉和图像处理领域已经表现出优越性，从而推进了人们对基于机器学习的行为识别进行越来越多的研究. 当前机器学习

的方法在行为识别领域所取得成就主要是针对RGB视频的研究. 这主要是因为设计一个3D输入的深度神经网络比较困难,许多机器学习方法仅仅将深度学习作为一个降维的方式.

Wu等[53]首先人工地从关节点提取EigenJoints[31]特征,然后再使用深度信念网络(deep belief network,DBN)来预测概率分布并进行分类. 而Liu等[52]提出了基于3D的深度卷积神经网络(3D-based deep convolutional neural network,3D^2 CNN),直接从深度序列自动地学习时空特征(如图7所示). 该方法首先对深度序列进行预处理,标准化成固定大小的立方体作为网络的输入,然后执行2个卷积层,每个卷积层执行最大池化层,最后执行3个全连接层并用softmax层进行分类.

Du等[54]则是根据人体的物理结构将人体骨架分成5部分,并放入5个双向递归神经网络

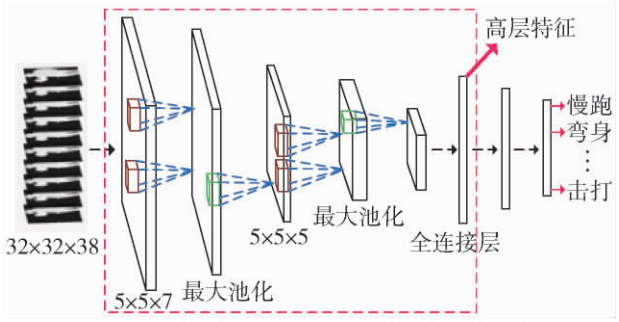


图7 3D深度卷积神经网络[52]

Fig. 7 3D deep convolutional neural network[52]

(bidirectional recurrent neural networks,BRNN)中. 随着层数的增加,分层地融合每一个子网络提取出的特征表示来构成更高层的输入. 最后使骨架序列的特征执行全连接层和softmax层,从而对动作进行分类(如图8所示).

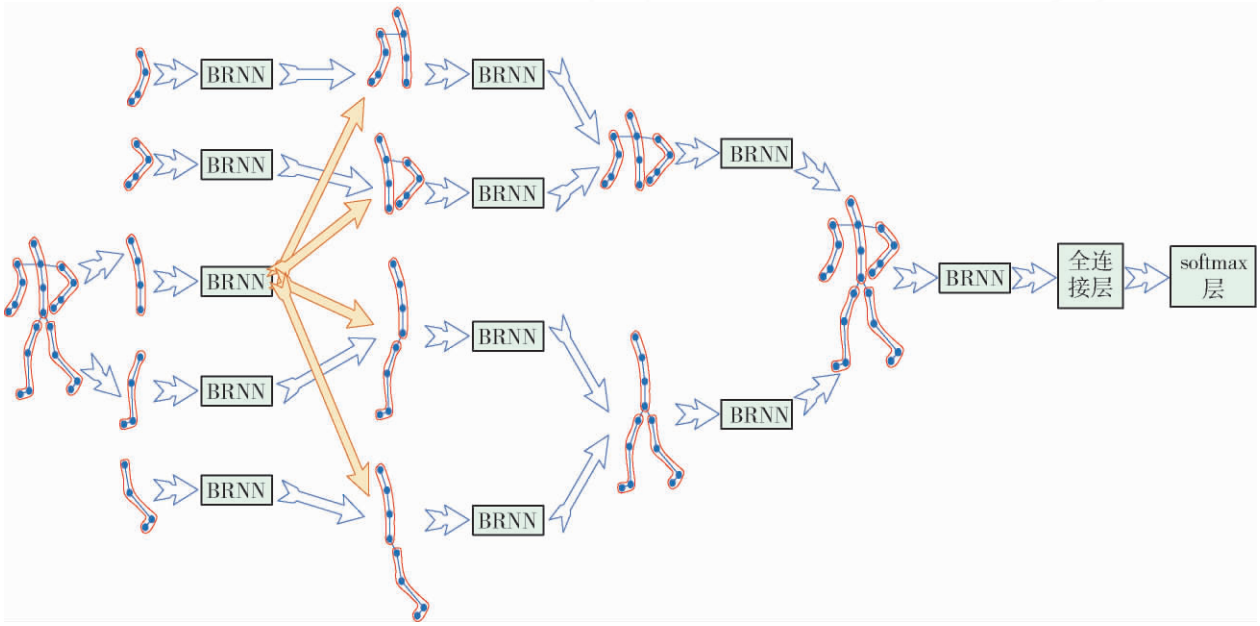


图8 分层双向递归神经网络的过程[54]

Fig. 8 Process of the hierarchical bidirectional recurrent neural network[54]

Veeriah等[55]提出了微分递归神经网络(differential recurrent neural network, dRNN)来自动学习具有显著时空序列的行为动态. 该方法将闸门机制(gating mechanism)加入到传统的长短时记忆(long short-term memory, LSTM)中来提取状态的导数(derivatives of state, DoS).

Liu等[56]认为2种特征进行融合不能充分地利用2个特征的互补性,提出了一种约束的基于图的遗传规划方法来自动地提取时空特征,同时融合了RGB和深度信息. 该方法通过输入层、过滤层和特征池化层获得特征向量,然后使用SVM分类.

本文介绍的提取特征的方法、所使用的分类器以及使用的数据库的情况见表2.

2 常用的公开数据库

为了能够公平地评价不同识别方法的性能,现在已经有很多公开的数据库供大家使用. 表3总结了18个公开的数据库,提供了发表时间、类别数、演员数、样本数以及数据模态,此外还提供了每个数据库的链接,便于读者下载. 下面将介绍其中最为常用的3个测试数据库: MSR-DailyActivity3D数据库、UCF Kinect数据库和MSR-Action3D数据库.

表 2　部分文献的概况

Table 2　Overview of some methods

特征表示	分类器	年份	数据库
HOJ3D[29]	HMM	2012	MSR-Action3D[57]、UTKinect-Action[29]
MSHs[30]	ELM	2016	MSR-Action3D[57]、Florence 3D-Action[58]、UTKinect-Action[29]
EigenJoints[31]	NBNN	2014	UCF Kinect[59]、MSR-Action3D[57]、CAD-60[60]
Weighted graphs[34]	序列匹配	2016	UCF Kinect[59]、MSRC-12[61]
JSGK[32]	SVM	2016	MSR-Action3D[57]、UTKinect-Action[29]、Florence 3D-Action[58]
Position offset[37]	NBNN	2016	UCF Kinect[59]、MSRC-12[61]
MP[35]	KNN	2013	MSR-Action3D[57]、MSR-DailyActivity3D[42]
MRF[36]	投票	2016	MSR-Action3D[57]、MSR-DailyActivity3D[42]、Huawei/3DLife[62]
LDP[25]	SVM	2012	RGBD-HuDaAct[25]
DCSF[28]	SVM	2013	MSR-Action3D[57]、MSR-DailyActivity3D[42]
STIP + T-Shape + T-MBH[26]	SVM/随机森林	2015	MSR-Action3D[57]、UTKinect-Action[29]、MSR-DailyActivity3D[42]
HON4D[39]	SVM	2013	MSR-Actions3D[57]、MSR-DailyActivity3D[42]
SNV[40]	SVM	2014	MSRActionPairs3D[39]、MSR-Actions3D[57]、MSR-DailyActivity3D[42]
ARP[41]	SVM	2016	MSR-Actions3D[57]、MSR-DailyActivity3D[42]
JAS + HOG^2[48]	SVM	2013	MSR-Actions3D[57]、UCF Kinect[9]
LOP[42]	SVM	2012	MSR-Action3D[57]、MSR-DailyActivity3D[42]
ROP[43]	SVM	2012	MSR-Action3D[57]
STOP[44]	SVM	2014	MSR-Action3D[57]
STIPs + 关节点特征[23]	随机森林	2013	MSR-Action3D[57]、UTKinect-Action[29]、CAD-60[60]
ScTPM + CS-Mltp[33]	SVM	2014	MSR-Action3D[57]、MSR-DailyActivity3D[42]
STIP + STACOG + EigenJoints + HON4D[49]	SVM/随机森林	2015	MSR-Action3D[57]、MSR-DailyActivity3D[42]、UTKinect-Action[29]、CAD-60[60]
GLAC + STACOG[51]	ELM	2017	MSR-Action3D[57]
$3D^2$CNN + JointVector[52]	SVM	2016	MSR-Action3D[57]、UTKinect-Action[29]
HBRNN[54]	RNN	2015	Berkeley MHAD[63]、MSR-Action3D[57]
RGGP[56]	SVM	2013	MSR-DailyActivity3D[42]

2.1　MSR-DailyActivity3D 数据库

MSR-DailyActivity3D 数据库[42]是通过 Kinect 采集日常活动中的 16 类人体动作而形成的一个数据库. 该数据库由 10 个演员完成，每个演员表演每个动作 2 次，总计含 320 个动作视频序列. 采集场景中有一个沙发，第 1 次是坐在沙发上完成动作，第 2 次是处于站立状态完成动作. 该数据库动作序列提供了相应动作的 RGB 视频序列以及每一动作帧的深度图像和关节点的位置信息. MSR-DailyActivity3D 数据库包含很多人与物交互的动作类，从而给识别算法造成困难，后文可见各算法在该数据库上的识别性能相比其他数据库均较差.

2.2　UCF Kinect 数据库

UCF Kinect 数据库[59]也同样包含 16 个类别的动作. 该数据库是通过 Kinect 传感器和 OpenNI 平台采集的，由 16 个演员执行，包括 13 个男士和 3 个女士，年龄段为 25 ~ 35 岁. 每类动作由每个演员执行 5 次，共采集 1 280 段视频序列. 数据记录形式为每一动作帧的 15 个人体关节点的 3D 位置坐标，但没有保存相应动作的 RGB 序列和深度图像. 对于该数据库，所使用的验证方式有很多，主要采用的是 n 折交叉验证的方式.

2.3　MSR-Action3D 数据库

MSR-Action3D 数据库[57]是一个建立比较早的数据库，大部分识别方法都是使用该数据库来评价其性能的. 该数据库包括 10 个演员的 20 个动作类. 每个演员在进行表演时都面对相机，每个动作完成 3 次，总计含 600 段动作视频序列. 该数据库

表 3 公开的数据库

Table 3 Public datasets

名字	年份	类别数	演员数	样本数	数据模态	链接
MSR-Action3D[57]	2010	20	10	567	深度图像 + 关节点	http: // research. microsoft. com/en-us/um/people/zliu/actionrecorsrc/
CAD-60[60]	2011	12	4	60	RGB + 深度图像 + 关节点	http: // pr. cs. cornell. edu/humanactivities/data. php
RGBD-HuDaAct[25]	2011	12	30	1 189	RGB + 深度图像	http: // adsc. illinois. edu/sites/default/files/files/ADSC-RGBD-dataset-download-instructions. pdf
Florence 3D-Action[58]	2012	9	10	215	关节点	http: // www. micc. unifi. it/vim/datasets/3dactions/
MSR-DailyActivity3D[42]	2012	16	10	320	RGB + 深度图像 + 关节点	http: // research. microsoft. com/en-us/um/people/zliu/actionrecorsrc/
UTKinect-Action[29]	2012	10	10	200	RGB + 深度图像 + 关节点	http: // cvrc. ece. utexas. edu/KinectDatasets/humanDetection. html
G3D[64]	2012	20	10	213	RGB + 深度图像 + 关节点	http: // dipersec. king. ac. uk/G3D/
MSRC-12[65]	2012	12	30	594	关节点	http: // research. microsoft. com/en-us/um/cambridge/projects/msrc12/
CAD-120[66]	2013	10	4	120	RGB + 深度图像 + 关节点	http: // pr. cs. cornell. edu/humanactivities/data. php
MSR Action Pairs[39]	2013	6	10	180	深度图像	http: // www. cs. ucf. edu/ ~ oreifej/
WorkoutSU-10 dataset[67]	2013	10	12	1 200	RGB + 深度图像 + 关节点	http: // vpa. sabanciuniv. edu. tr/phpBB2/vpa_views. php? s = 31&serial = 36
UCF Kinect[59]	2013	16	16	1 280	关节点	http: // www. cs. ucf. edu/ ~ smasood/datasets/UCFKinect. zip
Berkeley MHAD[63]	2013	11	12	659	RGB + 深度图像 + 关节点	http: // tele-immersion. citris-uc. org/berkeley_mhad
Huawei/3DLife[62]	2013	22	17	3 740	RGB + 深度图像	http: // mmv. eecs. qmul. ac. uk/mmgc2013/
MSR 3D online action[68]	2014	7	24	336	RGB + 深度图像 + 关节点	http: // research. microsoft. com/en-us/um/people/zliu/actionrecorsrc/
UWA3D Multiview[69]	2014	30	10	720	RGB + 深度图像 + 关节点	http: // staffhome. ecm. uwa. edu. au/ ~ 00053650/databases. html
UTD-MHAD[70]	2015	27	8	861	RGB + 深度图像 + 关节点	http: // www. utdallas. edu/ ~ kehtar/UTD-MHAD. html
NTU RGB + D[71]	2016	60	40	56 880	RGB + 深度图像 + 关节点	http: // rose1. ntu. edu. sg/datasets/actionrecognition. asp

提供了各动作的深度图像序列和 20 个人体关节点的 3D 位置信息. 这个数据库深度序列的背景相对比较干净,但是数据库中的一些动作类彼此之间非常相似,对识别算法性能要求较高. 表 4 列出了该数据库根据不同动作类划分的 3 个分类实验用子数据库. Action set 1(AS1) 和 Action set 2(AS2) 是将一些相似的动作类放在一起以供识别,而 Action set 3(AS3) 是将一些复杂的动作类放在一起以供识别.

比如，AS1 中的动作类 hammer 就特别容易与动作类 forward punch 混淆，而动作类 pickup throw 是 2 个动作类 bend 和 high throw 组成的. 不同的方法在该数据库上所采用的验证策略不同，大部分方法都会考虑数据质量、数量而选择演员 1、3、5、7、9 对应的序列进行训练，以其他演员完成的动作序列作为测试序列.

表 4　MSR-Action3D dataset 的 3 个子动作库

Table 4　Three subsets of MSR-Action3D dataset

AS1	AS2	AS3
Horizontal wave	High wave	High throw
Hammer	Hand Catch	Forward Kick
Forward punch	Draw X	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Hands wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup throw	Side boxing	Pickup throw

3　各方法识别性能的比较

用于度量各算法识别性能的最常用手段是计算该算法在指定测试数据库中的平均识别率，即可正确分类的序列数量相对该数据库序列总数的占比. 使用较多的测试数据库包括 MSR-DailyActivity3D 数据库、UCF Kinect 数据库和 MSR-Action3D 数据库. 表 5 收集了各方法在 MSR-DailyActivity3D 数据库、UCF Kinect 数据库和 MSR-Action3D 数据库(分别简写为 MSRDA、UCF 和 MSRA3D) 上的识别率. 其中，MSRA3D 对应的识别结果是对所有动作类的总体平均识别率，该数据库包含的 3 个子数据库所对应的识别率如表 6 所示. 对于这 2 个表，本文将同一特征表示类型的识别方法放在一起，可以直观地看出在每个类型中比较好的识别方法.

3.1　不同数据库上的算法性能对比

从表 5 中可以看出，各不同特征模型及不同分类器的识别算法在 3 个数据库上的性能表现的总体趋势是一致的，即从 UCF 到 MSRA3D 再到 MSRDA，算法识别性能递减.

在 UCF 上的算法识别率最高，普遍达到 96% 以上，文献 [34] 中的识别率最高，甚至达到 99.3% [34]. 分析其原因，除该数据库数据质量较好之外，还在于该数据库中的各动作类采集自 16 个演员的 5 次重复表演，即每个动作对应了 80 个数据样本，相比于其他 2 组数据库而言，具有更高密度的训练数据采样率，这必然有益于识别算法性能的提高.

在 MSRDA 上的总体识别率最低，仅文献 [33] 中所使用的多特征融合算法的识别率达到 92.5%，而除此之外的使用单个特征表示模型的算法识别率基本上均低于 90%. 分析其原因，一方面在于各动作数据样本个数少，仅有 20 个; 另一方面还主要在于该数据库所采集的人体日常行为数据相比于其他数据库更为复杂. 其大部分动作都涉及人-物之间的交互行为，存在较多的视觉遮挡，使得相应动作的人体关节点检测失效，对应特征模型的表达能力降低. 虽然增加了 UCF 所不具有的 RGB 形式的动作信息，仍不能弥补所提取特征的失效.

在 MSRA3D 上的识别率相比上述 2 个数据库，表现出与动作采样率一致的现象: 居二者之间且大部分识别方法的性能相差不大. 但是，因为该数据库所采集的某些行为的深度图像噪声很大，关节点的位置偏离得比较远，所以总体识别率仍偏低. 识别率最高的算法是文献 [33] 所提出的 93.83%.

由此可以看出，数据质量及数量的差别可直接导致基于训练的模型分类性能的差异，这已是模式识别领域的共识性结论.

3.2　特征模型对算法性能的影响分析

大部分识别方法都会用到 MSRA3D，因此，重点介绍该数据库有代表性的方法的性能. 被引用较多的方法是文献 [29,31,42,57]. 文献 [29] 提出了 3D 关节点位置直方图的方法; 文献 [31] 通过关节点之间的位置关系来描述一个动作序列; 文献 [42] 提出了一种局部占有模式; 文献 [57] 提出了一组 3D 点云来描述一个姿态，并构造了一个动作图. 通过表 5 对比文献 [29,31,42]，可以看出文献 [42] 对所有动作类的总体平均识别率表现出更好的综合性能，识别率达到 88.2%. 通过表 6 对比文献 [29,31,57] 可以看出: 文献 [29] 在 AS1 和 AS2 这 2 个子数据库表现出更好的性能，识别率分别达到 87.98% 和 85.48%，说明该方法更适合识别动作比较相似的行为; 文献 [31] 在 AS3 上表现出明显的优势，识别率达到 96.4%，说明该方法更适合识别一些比较复杂的行为; 文献 [57] 在这 3 个子数据库上的表现相对比较稳定.

在实验中，由于不同的方法使用不同的数据库和验证方案，所以很难说明哪一种方法是最好的方法. 每个数据库也呈现不同的特征，因此，每一种方

表 5 不同方法在 3 个数据库上的识别率

Table 5 Accuracies of different methods on three datasets

方法类别	文献	年份	分类器	识别率/%		
				MSRDA	UCF	MSRA3D
点特征	[28]	2013	SVM	88.20		89.30
	[26]	2015	随机森林	83.80		
	[29]	2012	HMM			78.97
	[30]	2016	ELM			90.98
线特征	[35]	2013	KNN	73.80		91.70
	[36]	2016	投票	78.52		91.01
	[33]	2014	SVM	90.63		93.83
	[72]	2015	KNN		99.20	92.10
	[59]	2013	逻辑回归		95.94	65.70
	[31]	2014	NBNN		97.10	82.30
	[37]	2016	NBNN		97.58	
	[73]	2015	HMM		97.66	89.23
	[74]	2014	HMM		98.90	
	[34]	2016	序列匹配		99.30	
面特征	[75]	2015	SVM		97.91	91.21
	[39]	2013	SVM			88.89
	[40]	2014	SVM	86.25		93.09
	[41]	2016	SVM	85.00		92.00
体特征	[42]	2012	SVM	85.75		88.20
	[43]	2012	SVM			86.20
	[44]	2014	SVM			81.55
融合	[49]	2015	随机森林	88.80		
	[33]	2014	SVM	92.50		
	[48]	2013	SVM		97.07	83.53
机器学习	[54]	2015	RNN			94.49
	[55]	2015	RNN			92.03

法在处理特定的数据库时都有自己的优点.

对于基于点特征的方法,一般会提取时空兴趣点,然后提出一种描述子进行特征表示. 最具竞争力的方法应该是文献[28],在 MSRDA 和 MSRA3D 上都表现出很好的性能,识别率分别达到 88.2% 和 89.3%. 文献[28]能够优于其他方法,是因为该方法提取时空兴趣点时采用了一种滤波方法,从而有效地抑制噪声,并使用了深度立方体相似特征描述局部 3D 深度立方体. 一般来说,这一类方法的性能要比其他类别的特征模型方法差,但是具有很强的适应性,所提出的描述子可用于识别人体或动物的行为,不依赖于骨骼信息或预处理,如人体检测、运动分割、跟踪以及图像去噪.

对于基于线特征的方法,一般会通过计算关节点之间的位置关系来表示特征. 这一类方法的性能相当,文献[33]相对更优,在 MSRDA 和 MSRA3D 两个数据库上都是自定义特征模型中识别率最高的,分别达到 90.63% 和 93.83%. 文献[33]首先计

表 6　不同方法在 MSR-Action3D 数据库上 3 个子数据集的识别率

Table 6　Accuracies of different methods on three subsets of MSR-Action3D dataset

方法类别	方法	年份	分类器	识别率/%		
				AS1	AS2	AS3
点特征	[57]	2010	GMM	72.90	71.90	79.20
	[29]	2012	HMM	87.98	85.48	63.46
	[30]	2016	ELM	89.76	91.70	92.50
线特征	[31]	2014	NBNN	74.50	76.10	96.40
	[73]	2015	HMM	90.29	95.15	93.29
体特征	[44]	2014	SVM	91.70	72.20	98.60
融合	[33]	2014	SVM	96.10	90.80	98.33
	[52]	2016	SVM	86.79	76.11	89.29
机器学习	[54]	2015	RNN	93.33	94.64	95.50

算了每一帧关节点之间的位置关系，然后为了对齐不同长度的序列，使用了基于稀疏编码的时域金字塔匹配的方法进行特征表示. 特征的线性和稀疏的组合很大程度上减少了提取特征的近似误差. 一般来说，这一类方法在保持性能较优的情况下，所提特征的复杂度比较低，权衡性能与复杂度也是目前识别算法的的主流趋势，因此，基于线特征的方法是很有前景的，即便还需要在其他数据库上更多地验证这种猜测.

基于面特征的方法，一般会通过计算曲面的法向量来提取特征. 这一类方法的性能相当，较好的方法是文献[40]，在 MSRDA 和 MSRA3D 两个数据库上识别率分别达到 86.25% 和 93.09%. 该方法聚集了超曲面法线来描述局部运动和形状信息，不仅计算了曲面法线，还保持了相邻法线的关系，使它们对噪声有更好的鲁棒性. 一般来说，这一类方法的性能要优于其他类别的方法，每一种方法都保持较高的识别率，不过这一类方法的复杂度相对也比较高.

基于体特征的方法，一般会提出一种占有模式(occupancy pattern)来提取特征. 比较好的方法是文献[42]，在 MSRDA 和 MSRA3D 两个数据库上识别率分别达到 85.75% 和 88.2%. 该方法使用了基于局部占有模式的方法，并使用傅里叶时域金字塔来表示特征. 这一类方法的性能和基于点特征方法的性能相当. 基于点特征的方法提取时空兴趣点时包括背景信息，基于体特征的方法更稀疏，只有特定点的信息. 因为在某些情况下，背景是有帮助的，而在另外一些情况下则是会带来干扰，所以很难评价哪类提取特征的方法更好.

对于基于机器学习的方法，一般选择 RNN 进行识别，都表现出不错的性能，在 MSRA3D 数据库上的识别率达到92%以上. 大部分方法会选择单层的 RNN，性能更优的文献[54]则是将人体分成 5 个部分分别通过 RNN 提取特征，然后再进行融合. 这一类方法是很有前景的方法，因为该类方法与其他的手动提取特征的方法相比非常有竞争力，也是当前的研究热点.

分析表明，对自定义特征进行点、线、面、体的特征级别分类时，基于线级别特征的识别算法总体表现优于其他级别的特征模型，这在某种程度上提示了人体运动特征的优势特征提取层次. 同时，结合基于机器学习的特征模型相比自定义特征模型在识别上的性能优势，在特征学习过程中考虑基于线特征强化约束的优化策略将有望得到更好的优选特征.

3.3　特征分类器对算法性能的影响分析

通过表 2、6、7 可以看出，大部分方法的特征分类器都会采用判别方法进行分类，比如 SVM[25,28,39,42]. 只有少数方法采用生成模型，特别是采用 HMM[29,73-74]. 另外，文献[59]采用逻辑回归的方法，文献[49]采用随机森林. 也有一些方法应用了非参数的方法，比如 KNN[72]、NBNN[31] 等. 由此可见，针对不同的数据库以及不同的人体识别方法，判别方法要优于生成方法，特别是 SVM 使用的更多.

总之，训练数据对于识别算法性能的影响是确定且多方面的. 一方面，训练数据的质量与数量，直接影响了基于训练生成的特征模型的泛化能力；另一方面，基于机器学习的识别算法的高性能表现，也印证了训练数据对于特征提取环节的影响相对分类器的无差别性. 后一方面的表现同时还说明，对于

人体运动这种复杂数据的识别而言,基于学习的特征模型比自定义特征模型具有识别优势. 而特征分类器的表现与特征模型的选择有关,当 RNN 将二者结合为一体时,可取得最高识别性能.

4 人体行为识别技术发展趋势分析及展望

人体行为识别研究所涉及的技术环节众多,应用领域宽泛,其发展方向一方面受到相关技术进步的推进,比如深度学习方法的出现所带来的识别模型及分类算法的变革,同时,也要面临不断推陈出新的应用需求所提出的新问题,比如视角无关的识别需求所提出的模型迁移的问题等. 两方面合力的存在,使得人体行为识别成为模式识别领域中发展方向、研究热点最具多样性的一个问题.

4.1 人体行为数据采集的大数据化趋势

针对现有算法的分析表明,训练数据质量与数量对识别算法的性能有着至关重要的影响,特别是深度学习技术的出现及应用,进一步强化了算法对数据的依赖性,这对数据的多源、高质量、大规模、多形式采集提出要求,形成了人体行为数据的大数据化采集需求. 此外,新的人体运动感知设备的出现,使得人体运动数据的多源、多样化采集成为可能,数据形式涵盖了从二维空间的 RGB 图像到三维时空的图像序列,再到四维时空的 RGBD 图像序列,乃至综合考虑视角、光照等采集条件的更高维的数据形式. 以上两方面的共同作用,使得大数据化的人体运动数据的采集以及训练数据的自动标注成为该领域研究趋势之一.

1) 人体行为数据的大数据化采集需求

目前,缺乏大规模的深度图像的数据库,每个数据库的动作类也不相同,因此,很难评价不同方法的性能. 分析目前数据库的局限性,一方面在于现有的数据库只有少量的动作类,所提的方法只可以对某些数据库表现出较好的性能. 如果未来的数据库能够增加动作类的数量,那么将会对识别方法带来新的挑战,也可以更加公平地比较不同方法的性能;另一方面在于现有的数据库样本数量有限,使得一些数据驱动的学习方法不能应用到这个问题上,特别是最近比较热门的深度学习.

2) 人体行为数据的多样化采集需求

现有的数据库大部分都比较单一,目前的识别方法虽然在数据库上有很高的性能,但还很难应用到现实场景中. 目前数据库的局限性如下:首先,现有数据库所采用的相机的视角非常受限,大部分数据库的样本都是从一个固定的视角来采集的,还有一部分数据库则是同时使用多个相机,但也是固定的前视角和侧视角. 其次,现有的数据库表演者年龄范围比较窄,类内的差异很有限. 人体行为的构成取决于人的年龄、性别、文化甚至是身体状况,因此,表演者的差异对行为识别的数据库是很重要的. 然后,现有的数据库大部分是在室内场景下采集的,数据库的背景都只有一种颜色,或者只在一个固定不变的场景采集. 最后,现有的数据库大部分是在理想的环境下采集的,没有考虑到光照变化的影响,也没有足够多的遮挡情况.

4.2 特征表示模型性能与识别算法效率并重的趋势

针对识别算法所做的性能分析表明,多特征融合以及基于学习特征的表示模型取得比自定义特征表示模型更优的识别效果. 但是,在构建更为复杂有效的特征表示模型以实现更高准确率的动作识别的同时,也将会不可避免地遭遇算法运行效率降低的困境,二者的折衷是目前算法的必然之选. 但从发展的角度来看,性能与效率同步提高的识别算法的提出是顺应技术进步的主流趋势. 具体可表现为低延时的高性能算法设计以及基于融合特征模型的高效率识别算法设计.

1) 低延时的高性能行为识别

现有识别方法的识别率在提高的同时,计算复杂度也会增加,从而会带来高延时. 但是在实际应用中,低延时识别是一个具有重要研究价值的问题,特别是人机交互对实时性的要求很高,因此,权衡识别率和低延时是一个极大的挑战. 从当前的研究发现,基于 3D 人体关节点的方法可以获得与使用 RGBD 多模态数据的现有方法相近甚至更高的识别率,具有很强的判别性,并且所需要的计算量和存储信息都很少,因此,基于人体关节点的行为识别为研究这一问题寻找到了方向. 但是现有的低延时的行为识别方法还不适合应用到现实场景中,这是因为许多方法不可以处理无分割的行为序列,这对现实场景是很重要的;目前缺少大量现实的行为数据库,许多当前的方法仅仅局限于针对少量动作类别;实际应用中不可避免地会遇到遮挡问题,当前的方法并不能很好地解决遮挡问题,并且目前存在的数据库也没有包括足够多的遮挡情况. 所以这些部分的改善将会使低延时行为识别方法更好地应用到实际应用中.

2) 基于融合特征模型的高效率行为识别

不同模态下提取特征的方法不同,相同模态下

提取的特征是从不同的角度对人体行为进行描述，这些提取特征的方法相互之间具有很强的互补性. 尽管现在已经提出了很多融合的方法，但是有些方面还需要深入研究. 比如：不同特征在识别不同的动作类时，识别率会有很大的不同，不同特征的融合可能是解决复杂人体行为识别问题的一种思路；特征之间最优的权重分配问题是一个尚未解决的问题，当前方法的权重分配一般都是由训练样本决定的，对所有的测试样本都是固定的，因此，根据每个测试样本特征的自适应权重分配是一个值得研究的问题；当前的融合方法虽然识别率上有一定的提高，但是所提出算法的复杂度也在增加，计算时间也随之增加，这样也就缺少了实用价值. 因此，基于融合的实时识别的方法也是一个需要探讨的问题.

4.3　面向应用的行为识别算法适用空间日益扩展的趋势

为了顺应行为识别应用领域的不断扩展，相关识别技术的研究范畴及适用范围也日益扩展，主要表现为对识别算法能够适用的空间条件正逐渐从满足理想、半理想假设的环境，向现实环境不断地迁移和扩展. 在可预见的将来，视角无关的动作识别研究以及大范围监控环境下的人体行为识别将成为重要的研究课题.

1）视角无关的识别方法

在实际应用中，因为视频都是在任意视点获取的，所以在实际应用中的识别方法就必须要与视角无关. 目前，通过传感器可以获取场景中更多的结构信息，视角无关识别方法相比以前也更容易实现，然而，基于深度图像的视角无关的分析和解决方案依然比较缺乏. 大多数方法主要是基于单一视角，并且大部分数据库也是在单一视角下完成的. 因此，未来需要研究复杂的视角无关的识别方法来应对动作的多样性、视角的随意性等问题，还需要采集大量的多视角的数据库为不同识别方法提供一个公平的评价平台.

2）大范围监控环境下的人体行为识别

目前的方法主要是针对单个人的识别，大范围监控环境下多个人体目标的群体行为识别是一个极具挑战性的问题. 主要是由于当前的深度相机探测范围的限制，现在提出的人体行为识别方法仅限于在较小范围的环境下使用，还不能应用到大范围监控环境下. 但是，随着传感器技术的不断发展，远距离的深度传感器可能会问世，群体的深度图像也将会成为可能，大范围监控环境下的群体行为识别方法将会成为一个研究课题.

总之，人体行为识别是一个既复杂又非常重要的研究方向. 提高人体行为识别的性能，不仅需要考虑模型本身，还需要考虑多样化的大规模的公开数据库. 这些方面的改善可以极大地提高该技术在现实中的应用，特别是游戏应用和人机交互.

参考文献：

[1] CHEN H Y, HWANG J N. Integrated video object tracking with applications in trajectory-based event detection [J]. Journal of Visual Communication and Image Representation, 2011, 22(7): 673-685.

[2] PIRSIAVASH H, RAMANAN D. Detecting activities of daily living in first-person camera views [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2012: 2847-2854.

[3] ROCCETTI M, MARFIA G, SEMERARO A. Playing into the wild: a gesture-based interface for gaming in public spaces [J]. Journal of Visual Communication and Image Representation, 2012, 23(3): 426-440.

[4] YU G, YUAN J S, LIU Z C. Real-time human action search using random forest based hough voting [C]// Proceedings of the 19th ACM International Conference on Multimedia. New York: ACM, 2011: 1149-1152.

[5] BIAN W, TAO D, RUI Y. Cross-domain human action recognition [J]. IEEE Transactions on Systems Man & Cybernetics: Part B, 2012, 42(2): 298-307.

[6] ZHANG Z, TAO D. Slow feature analysis for human action recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(3): 436-450.

[7] MATIKAINEN P, HEBERT M, SUKTHANKAR R. Trajectons: action recognition through the motion analysis of tracked features [C]// Proceedings of IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Piscataway: IEEE, 2009: 514-521.

[8] POPPE R. A survey on vision-based human action recognition [J]. Image and Vision Computing, 2010, 28(6): 976-990.

[9] POPPE R. Vision-based human motion analysis: an overview [J]. Computer Vision & Image Understanding, 2007, 108(1): 4-18.

[10] TURAGA P, CHELLAPPA R, SUBRAHMANIAN V S, et al. Machine recognition of human activities: a survey [J]. IEEE Transactions on Circuits & Systems for Video Technology, 2008, 18(11): 1473-1488.

[11] BOBICK A F, DAVIS J W. The recognition of human

movement using temporal templates [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001, 23(3): 257-267.

[12] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2008: 1-8.

[13] LIU H, SUN M T, WU R C, et al. Automatic video activity detection using compressed domain motion trajectories for H. 264 videos [J]. Journal of Visual Communication and Image Representation, 2010, 22(5): 432-439.

[14] SUN J, WU X, YAN S, et al. Hierarchical spatio-temporal context modeling for action recognition [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2009: 2004-2011.

[15] SHOTTON J, FITZGIBBON A, COOK M, et al. Real-time human pose recognition in parts from single depth images [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2013: 1297-1304.

[16] SHABANI A H, CLAUSI D A, ZELEK J S. Evaluation of local spatio-temporal salient feature detectors for human action recognition [C]// Proceedings of the 9th Conference on Computer and Robot Vision. Piscataway: IEEE, 2012: 468-475.

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

[18] WILLEMS G, TUYTELAARS T, GOOL L V. An efficient dense and scale-invariant spatio-temporal interest point detector[C]// Proceedings of European Conference on Computer Vision(ECCV). Berlin: Springer, 2008: 650-663.

[19] DOLLARP, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatio-temporal features [C]// Proceedings of the 14th International Conference on Computer Communications and Networks. Piscataway: IEEE, 2005: 65-72.

[20] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2005: 886-893.

[21] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[22] BO L, LAI K, REN X, et al. Object recognition with hierarchical kernel descriptors[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2011: 1729-1736.

[23] ZHU Y, CHEN W, GUO G. Fusing spatiotemporal features and joints for 3D action recognition [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2013: 486-491.

[24] NI B, WANG G, MOULIN P. RGBD-HuDaAct: a color-depth video database for human daily activity recognition [C]// IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Piscataway: IEEE, 2011: 1147-1153.

[25] ZHAO Y, LIU Z, YANG L, et al. Combing RGB and depth map features for human activity recognition [C]// Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway: IEEE, 2012: 1-4.

[26] CHEN W, GUO G. 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.

[27] CHENG Z, QIN L, YE Y, et al. Human daily action analysis with multi-view and color-depth data [C]// Proceedings of IEEE International Conference on Computer Vision (ICCV). Berlin: Springer, 2012: 52-61.

[28] XIA L, AGGARWAL J K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2013: 2834-2841.

[29] XIA L, CHEN C C, AGGARWAL J K. View invariant human action recognition using histograms of 3D joints [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2012: 20-27.

[30] SALIH A A A, YOUSSEF C. Spatiotemporal representation of 3D skeleton joints-based action recognition using modified spherical harmonics [J]. Pattern Recognition Letters, 2016, 83: 32-41.

[31] YANG X, TIAN Y L. Effective 3D action recognition using EigenJoints [J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 2-11.

[32] LI M, LEUNG H. Graph-based approach for 3D human skeletal action recognition [J]. Pattern Recognition Letters, 2016, 87: 195-202.

[33] LUO J J, WANG W, Qi H R. Spatio-temporal feature extraction and representation for RGB-D human action recognition[J]. Pattern Recognition Letters, 2014, 50: 139-148.

[34] JIANG X B, ZHONG F, PENG Q S, et al. Action recognition based on global optimal similarity measuring [J]. Multimedia Tools and Applications, 2016, 75 (18): 11019-11036.

[35] ZANFIR M, LEORDEANU M, SMINCHISESCU C. The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection [C] // Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2013: 2752-2759.

[36] CAI X Y, ZHOU W G, WU L, et al. Effective active skeleton representation for low latency human action recognition [J]. IEEE Transactions on Multimedia, 2016, 18(2): 141-154.

[37] LU G L, ZHOU Y Q, LI X Y, et al. Efficient action recognition via local position offset of 3D skeletal body joints[J]. Multimedia Tools and Applications, 2016, 75 (6): 3479-3494.

[38] TANG S, WANG X, LÜ X, et al. Histogram of oriented normal vectors for object recognition with a depth sensor [C] // Asian Conference on Computer Vision. Berlin: Springer, 2012: 525-538.

[39] OREIFEJ O, LIU Z. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2013: 716-723.

[40] YANG X D, TIAN Y L. Super normal vector for activity recognition using depth sequences [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2014: 804-811.

[41] ZHOU Y, MING A. Human action recognition with skeleton induced discriminative approximate rigid part model[J]. Pattern Recognition Letters, 2016, 83: 261-267.

[42] WANG J, LIU Z C, WU Y, et al. Mining actionlet ensemble for action recognition with depth cameras[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2012: 1290-1297.

[43] WANG J, LIU Z, CHOROWSKI J, et al. Robust 3D action recognition with random occupancy patterns[C] // Proceedings of European Conference on Computer Vision (ECCV). Berlin: Springer, 2012: 872-885.

[44] VIEIRA A W, NASCIMENTO E R, OLIVEIRA G L, et al. On the improvement of human action recognition from depth map sequences using Space-Time Occupancy Patterns [J]. Pattern Recognition Letters, 2014, 36: 221-227.

[45] LLINAS J, HALL D L. An introduction to multisensor data fusion [C] // IEEE International Symposium on Circuits & Systems. Piscataway: IEEE, 1998: 537-540.

[46] ATREY P K, HOSSAIN M A, SADDIK A E, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia Systems, 2010, 16(6): 345-379.

[47] CHANG K I, BOWYER K W, FLYNN P J. Face recognition using 2D and 3D facial data [C] // ACM Workshop in Multimodal User Authentication. New York, NY: ACM, 2004: 25-32.

[48] OHN-BAR E, TRIVEDI M M. Joint angles similarities and HOG^2 for action recognition [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2013: 465-470.

[49] ZHU Y, CHEN W, GUO G. Fusing multiple features for depth-based action recognition[J]. ACM Transactions on Intelligent Systems & Technology, 2015, 6(2): 1-20.

[50] KOBAYASHI T, OTSU N. Motion recognition using local auto-correlation of space-time gradients [J]. Pattern Recognition Letters, 2012, 33(9): 1188-1195.

[51] CHEN C, ZHANG B C, HOU Z J, et al. Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features [J]. Multimedia Tools & Applications, 2017, 76(3): 4651-4669.

[52] LIU Z, ZHANG C, TIAN Y. 3D-based deep convolutional neural network for action recognition with depth sequences[J]. Image & Vision Computing, 2016, 55: 93-100.

[53] WU D, SHAO L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2014: 724-731.

[54] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 1110-1118.

[55] VEERIAH V, ZHUANG N, QI G J. Differential recurrent neural networks for action recognition [C] //

Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 4041-4049.

[56] LIU L, SHAO L. Learning discriminative representations from RGB-D video data [C] // Proceedings of International Joint Conference on Artificial Intelligence. New York: ACM, 2013: 1493-1500.

[57] LI W Q, ZHANG Z Y, LIU Z C. Action recognition based on a bag of 3D points [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2010: 9-14.

[58] SEIDENARI L, VARANO V, BERRETTI S, et al. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2013: 479-485.

[59] ELLIS C, MASOOD S Z, TAPPEN M 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.

[60] SUNG J, PONCE C, SELMAN B, et al. Unstructured human activity detection from RGBD images [C] // Proceedings of IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2011: 842-849.

[61] FOTHERGILL S, MENTIS H, KOHLI P, et al. Instructing people for training gestural interactive systems [C] // Proceedings of Sigchi Conference on Human Factors in Computing Systems. New York: ACM, 2012: 1737-1746.

[62] Huawei. Huawei/3DLife ACM Multimedia Grand Challenge for 2013 [DB/OL]. [2017-04-18]. http: // mmv. eecs. qmul. ac. uk/mmgc2013/.

[63] VIDAL R, BAJCSY R, OFLI F, et al. Berkeley MHAD: a comprehensive multimodal human action database [C] // IEEE Workshop on Applications of Computer Vision. Piscataway: IEEE, 2013: 53-60.

[64] BLOOM V, MAKRIS D, ARGYRIOU V. G3D: a gaming action dataset and real time action recognition evaluation framework [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2012: 7-12.

[65] FOTHERGILL S, MENTIS H, KOHLI P, et al. Instructing people for training gestural interactive systems [C] // Proceedings of Sigchi Conference on Human Factors in Computing Systems. New York: ACM, 2012: 1737-1746.

[66] KOPPULA H S, GUPTA R, SAXENA A. Learning human activities and object affordances from RGB-D videos [J]. International Journal of Robotics Research, 2013, 32(8): 951-970.

[67] NEGIN F, ÖZDEMIR F, AKGÜL C B, et al. A decision forest based feature selection framework for action recognition from RGB-Depth cameras [C] // Proceedings of International Conference Image Analysis and Recognition. Berlin: Springer, 2013: 648-657.

[68] YU G, LIU Z C, YUAN J S. Discriminative orderlet mining for real-time recognition of human-object interaction [C] // Asian Conference on Computer Vision. Cham: Springer, 2014: 50-65.

[69] RAHMANI H, MAHMOOD A, DU Q H, et al. HOPC: histogram of oriented principal components of 3D pointclouds for action recognition [C] // Proceedings of European Conference on Computer Vision (ECCV). Cham: Springer, 2014: 742-757.

[70] CHEN C, JAFARI R, KEHTARNAVAZ N. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor [C] // IEEE International Conference on Image Processing. Piscataway: IEEE, 2015: 168-172.

[71] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB + D: a large scale dataset for 3D human activity analysis [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 1010-1019.

[72] DEVANNE M, WANNOUS H, BERRETTI S, et al. 3-D human action recognition by shape analysis of motion trajectories on riemannian manifold [J]. IEEE Transactions on Cybernetics, 2015, 45(7): 1340-1352.

[73] PRESTI L L, CASCIA M L, SCLAROFF S, et al. Hankelet-based dynamical systems modeling for 3D action recognition [J]. Image & Vision Computing, 2015, 44: 29-43.

[74] BEH J, HAN D K, DURASIWAMI R, et al. Hidden Markov model on a unit hypersphere space for gesture trajectory recognition [J]. Pattern Recognition Letters, 2014, 36: 144-153.

[75] SLAMA R, WANNOUS H, DAOUDI M, et al. Accurate 3D action recognition using learning on the Grassmann manifold [J]. Pattern Recognition, 2015, 48(2): 556-567.

(责任编辑 梁 洁)