RGB-D 行为识别研究进展及展望

胡建芳 [1, 2, 3] 王熊辉 [4] 郑伟诗 [1, 2, 3] 赖剑煌 [1, 2, 3]

摘 要 行为识别是计算机视觉领域很重要的一个研究问题, 其在安全监控、机器人设计、无人驾驶和智能家庭设计等方面都有着非常重要的应用. 基于传统 RGB 视频的行为识别方法由于容易受背景、光照等行为无关因素的影响, 导致识别精度不高. 廉价 RGB-D 摄像头出现之后, 人们开始从一个新的途径解决行为识别问题. 基于 RGB-D 摄像头的行为识别通过聚合 RGB、深度和骨架三种模态的行为数据, 可以融合不同模态的行为信息, 从而可以克服传统 RGB 视频行为识别的缺陷, 也因此成为近几年的一个研究热点. 本文系统地综述了 RGB-D 行为识别领域的研究进展和展望. 首先, 对近年来 RGB-D 行为识别领域中常用的公共数据集进行简要的介绍; 同时也系统地介绍了多模态 RGB-D 行为识别研究领域的典型模型和最新进展, 其中包括卷积神经网络 (Convolution neural network, CNN) 和循环神经网络 (Recurrent neural network, RNN) 等深度学习技术在 RGB-D 行为识别的应用; 最后, 在三个公共 RGB-D 行为数据库上对现有方法的优缺点进行了比较和分析, 并对未来的相关研究进行了展望.

关键词 RGB-D, 行为识别, 骨架点, 深度学习

引用格式 胡建芳, 王熊辉, 郑伟诗, 赖剑煌. RGB-D 行为识别研究进展及展望. 自动化学报, 2019, **45**(5): 829−840

DOI 10.16383/j.aas.c180436

RGB-D Action Recognition: Recent Advances and Future Perspectives

HU Jian-Fang[1, 2, 3] WANG Xiong-Hui[4] ZHENG Wei-Shi[1, 2, 3] LAI Jian-Huang[1, 2, 3]

Abstract Action recognition is an important research topic in computer vision, which is critical in some real-world applications including security monitoring, robot design, self driving and smart home system etc.. The existing single modality RGB based action recognition approaches are easily suffered from the illumination variation, background clutter, which leads to an inferior recognition performance. The emergence of low-cost RGB-D cameras opens a new dimension for addressing the problem of action recognition. It can overcome the drawbacks of single modality by outputting RGB, depth, and skeleton modalities, each of which can describe actions from one perspective. In this paper, we mainly review the current advances in RGB-D action recognition. Firstly, we briefly introduce some datasets popularly used in the research of RGB-D action recognition, then we review the literatures and the state-of-the-art recognition models based on convolution neural network (CNN) and recurrent neural network (RNN). Finally, we discuss the advantages and disadvantages of these methods through the experiments on three datasets and provide some problems needing addressing in the future.

Key words RGB-D, action recognition, skeleton, deep learning

Citation Hu Jian-Fang, Wang Xiong-Hui, Zheng Wei-Shi, Lai Jian-Huang. RGB-D action recognition: recent advances and future perspectives. *Acta Automatica Sinica*, 2019, **45**(5): 829−840

从图像视频中分析和理解人体行为是计算机视觉与模式识别领域的重要研究课题之一, 其在安全监控、机器人设计、无人驾驶和智能家庭设计等方面都有着非常重要的应用. 近年来, 由于图像视频拍摄设备制造技术的飞速发展, 人们可获得的视频图像语言越来越趋于多样化和复杂化, 其获得途径也越来越便捷化. 多模态视频图像同步记录设备的快速发展给相关的计算机智能应用技术, 特别是多媒体视频安全监控方面, 提供了新的发展契机, 一系列的基于多模态摄像头的研究课题和应用层出不穷. 特别是在廉价 RGB-D ("RGB-D" 指同时使用 RGB、深度和骨架三种模态数据) 摄像头出现之后, 人们开始尝试用一个新的途径 (深度信息) 来解决传统的计算机视觉、模式识别和计算机图形学问题[1−6].

收稿日期 2018-06-20 录用日期 2018-11-05
Manuscript received June 20, 2018; accepted November 5, 2018
国家自然科学基金 (61702567, 61876104), 广东省重大项目 (2018B010109007), 广东省信息安全技术重点实验室开放课题基金 (2017B030314131) 资助
Supported by National Natural Science Foundation of China (61702567, 61876104), Major Projects in Guangdong Province (2018B010109007), and the Opening Project of GuangDong Province Key Laboratory of Information Security Technology (2017B030314131)
本文责任编委 王亮
Recommended by Associate Editor WANG Liang
1. 中山大学数据科学与计算机学院 广州 510006 2. 广东省信息安全技术重点实验室 广州 510006 3. 机器智能与先进计算教育部重点实验室 广州 510006 4. 中山大学电子信息与工程学院 广州 510006
1. School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006 2. Guangdong Province Key Laboratory of Computational Science, Guangzhou 510006 3. Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006 4. School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006

与传统的 RGB 数据相比, 多模态的 RGB-D 数据可以给行为分析方面的研究带来不少便利. RGB 图像数据容易受拍摄环境, 光照和行为人衣着纹理等与行为无关的外界因素影响, 直接从 RGB 视频图像中推断行为人的骨架姿势、轮廓信息和一些关键动作信息是件很困难的事情, 从而导致很多视频分析和行为动作分析技术在实际生活中没有得到很好的应用[7]. 如图 1 所示, 在深度视频图像中, 因行人与周围的拍摄场景通常具有很高的辨识度, 且所获得的深度数据不容易受衣着的影响, 从中获得行人轮廓骨架信息简单方便准确很多; 而 RGB 视频中的颜色信息能更细致地刻画物体表观纹理特征, 这些在处理涉及人与物体交互的行为[1, 8] 时显得特别重要. 多模态 RGB-D 数据虽然可以为行为识别研究提供更多的信息, 但同时也给相关的视频分析研究带来了新的挑战. 首先, 不同模态数据从不同角度刻画行为信息, 传统 RGB 视频图像分析领域中的常用特征如 HOG (Histogram of oriented gradient)[9]、SIFT (Scale invariant feature transform)[10]、LBP (Local binary pattern)[11] 等并不一定适用于其他模态的视频图像数据, 怎样从深度摄像仪器拍摄的深度数据或者 3 维骨架数据中挖掘出有效动作变化信息进行行为表示及识别, 是该领域的一个研究难点. 其次, 多模态 RGB-D 摄像数据包含多个模态, 怎样才能更有效地合并不同模态的信息以获得更多的行为上下文内容信息 (Context) 使识别能达到更好的效果, 也是 RGB-D 行为识别的研究热点之一.

为了克服上述挑战, 已经提出了很多 RGB-D 行为识别算法, 它们在构建深度特征描述子、三维骨架动态特征提取、多模态特征融合等方面采用了不同的策略. 本文分别从数据、模型方法和实验结果分析三个方面比较系统全面地介绍了目前 RGB-D 行为识别研究现状. 在模型方法介绍方面, 本文按照模型所使用的数据模态对现有的方法进行了分类介绍, 并结合多个公共数据库中的实验结果分析了相关方法的优缺点.

1 RGB-D 行为公共数据库介绍

与其他数据驱动为主的视觉应用问题一样, 数据在行为识别中也起着非常重要的作用. 为了促进 RGB-D 行为识别方面的研究, 国内外研究者从不同研究角度收集了大量的 RGB-D 行为数据库, 不同的数据库包含了用 Kinect 拍摄

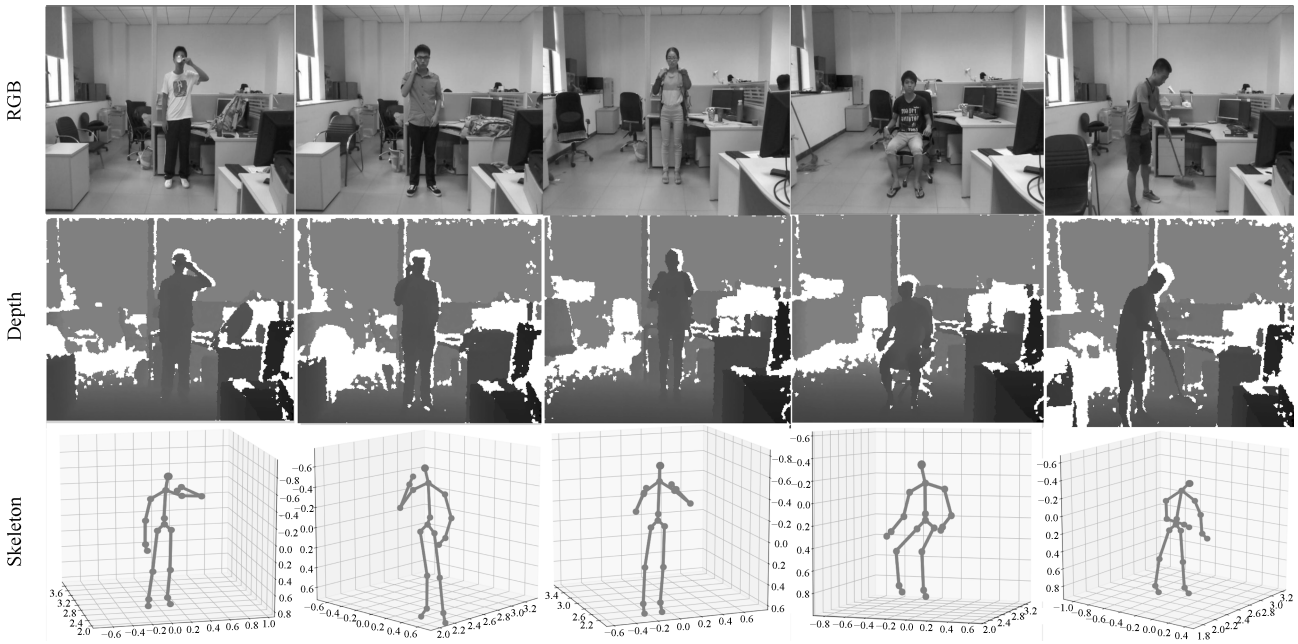


图 1 RGB-D 数据样例 (图中为 SYSU 3DHOI 数据库中的部分样本, 从上到下依次为彩色数据 (RGB), 深度数据 (Depth) 和骨架数据 (Skeleton), 从左到右的所对应的行为分别为 “喝水”、“打电话”、“背包”、“坐下”、“打扫”. 从图中可以看到, 每种模态的数据从不同角度刻画行为内容.)

Fig. 1 Some RGB-D samples captured by Kinect (This figure presents some samples from SYSU 3DHOI set. The examples for RGB, depth and skeleton modalities are provided in the first, second, and third rows, respectively. Each column in the figure gives a sample of action “drinking”, “calling”, “packing”, “sitting down”, and “sweeping”, respectively. As shown, each of the modalities characterizes actions from one perspective.)

的不同应用背景的行为视频数据. 公共数据库的发展在一定程度上反映了研究主流方法的发展. 在深度学习被大范围应用到 RGB-D 行为识别之前, 收集的 RGB-D 行为数据库都是相对比较小规模的, 其总样本数不超过 5 000, 行为类别数也不超过 20. 深度学习兴起之后, 大规模的 RGB-D 行为数据也开始出现, 以配合深度学习算法应用. 下面, 本文将对 RGB-D 行为识别研究领域中具有一定代表性的数据库进行简要介绍.

1.1 MSR 日常行为数据库

MSR 日常行为数据库[12] 是由 Wang 等在微软雷德蒙研究院所创建, 其包含 10 个行为人拍摄的 16 种日常行为视频 (如喝水、看书、鼓掌等), 每种行为都以站立和坐着的方式重复拍摄 2 次. 因此, 该库总共有 320 个视频, 每个视频都记录了相应的深度视频、RGB 视频和三维骨架序列数据. 特别地, 该库中的大部分行为都包含人与物体之间的交互动作. 为了测试模型性能, 数据库构建者采用了传统的个体交叉的验证方式进行实验验证, 即将其中 5 个行为人拍摄的 160 个行为视频用来训练模型, 剩下行为人相关的 160 个视频用来测试.

1.2 SYSU 3DHOI 行为数据库

SYSU 3DHOI[13] 是一个专门关注于人与物体交互行为的数据库. 为了搭建该数据库, 来自于中山大学的 Hu 等邀请了 40 位参与者尽可能自由地做 12 种不同的交互行为 (如喝水、倒水、打电话和玩手机等). 这 12 个交互行为主要涉及 6 种不同的被操作物体: 手机、椅子、书包、钱包、扫把和拖把, 每种物体都与其中 2 个不同的交互行为相关. 因此, 该数据库总共包含有 480 个 RGB-D 视频. 创建人设置了两种不同的测试方案. 第一种测试为: 从每个行为类中随机选取一半的视频作为训练集, 剩下的一半作为测试. 第二种测试为: 随机选取 20 个个体的视频数据作为训练, 剩下的个体视频进行测试. 第二种测试为传统的个体交叉认证. 上述的每种测试方案都重复进行 30 次后取平均结果作为最终识别效果.

1.3 多视角 3D 行为数据库

多视角 3D 行为数据库[14] 是由西安理工大学的 Wei 等于 2013 年等建立. 创建该库主要初衷是为了研究跨视角的 RGB-D 行为识别问题. 为了拍摄该库, 创建者邀请了 8 个个体实施预先定义好的 8 种交互行为 (用手机打电话、用杯子喝水、倒水、打水、按按钮、看书、用鼠标和敲键盘等), 每个行为重复拍摄大概 20 次左右. 所有个体的行为实施过程被三个 Kinect 摄像头从不同的角度同时捕捉拍摄. 该库是个比较大的规模的行为库, 其总共包含 3 815 个行为序列, 383 036 个 RGB-D 视频帧. 每个行为类别对应有 477 个左右的行为视频. 作者在其主页上公开了部分的行为视频数据.

1.4 CAD60 行为数据库

CAD60 行为数据库[15] 是由康奈尔大学的 Sung 等拍摄. 该库总共包含由 Kinect 拍摄的 68 个视频. 为了拍摄该数据库, Sung 等邀请了 4 个行为人分别进行 13 种特定的行为 (含静止站立、打电话等), 每个行为样本可能涉及如下 5 种场景之一: 办公室、厨房、卧室、洗浴间和客厅. 本数据库采用了针对每种场景的留一法交叉验证的方式对模型进行训练测试, 即对于一种特定场景, 其中三个行为个体的视频样本用来训练, 剩下的用来测试. 这样可以保证训练集和测试集中不会出现同一个人. 因此, 该库中总共涉及 20 次训练测试, 平均识别效果作为最终的识别结果. 后来, 该库被进一步拓展为 CAD120[16].

表 1 现有 RGB-D 行为数据库的对比 (更完整的数据库介绍请参见文献 [17])

Table 1 Comparison of some existing RGB-D action datasets (Please refer to [17] for more details about the datasets)

数据库	数据类型	类别数	个体数	视频数	交互比例	是否公开下载	发表年份
MSRAction[18]	Depth	20	10	567	$\leq 70\%$	全部公开	2010
CAD 60[15]	RGB-D	14	4	68	85.7 %	全部公开	2011
UTKinect[19]	RGB-D	10	10	200	$\geq 30\%$	全部公开	2012
MSRActionPair[20]	RGB-D	6	10	360	100 %	全部公开	2013
CAD-120[16]	RGB-D	10	4	120	100 %	全部公开	2013
MSRDaily[12]	RGB-D	16	10	320	87.5 %	全部公开	2013
Multiview[14]	RGB-D	8	8	3 815	100 %	部分公开	2013
RGBD-HuDaAct[21]	RGB-D	12	30	1 189	100 %	全部公开	2013
Comp. Activities[22]	RGB-D	16	14	693	75 %	全部公开	2014
ORGBD[23]	RGB-D	7	36	386	100 %	全部公开	2014
TJU dataset[24]	RGB-D	22	20	1 760	$\leq 13.6\%$	全部公开	2015
SYSU 3DHOI[1]	RGB-D	12	40	480	100 %	全部公开	2016
NTU[25]	RGB-D	60	40	56 880	100 %	全部公开	2016

1.5 NTU 大规模行为数据库

NTU 大规模数据库[25] 是目前包含行为样本数目最多的 RGB-D 数据库, 由来自于新加坡南洋理工大学的 Shahroudy 等于 2016 年创建. 该库由第二代 Kinect 拍摄, 因而其深度和彩色视频的分辨率比前面两个库大. 其包含来自于 60 个行为类 3 种不同视觉下的 56 880 个 RGB-D 视频. 与其他数据库相比, 该库中考虑的行为更为复杂, 它们可能包含个体的手势动作 (如跳跃、鼓掌等), 人与物体交互的行为 (如喝水、吃零食等) 和人与人交互的行为 (如拥抱、用手指指着别人等) . 为了实验测试, 作者设置了两种不同的训练测试集划分: 个体交叉和视角交叉. 在个体交叉中, 20 个个体的行为数据被用来作为训练集, 剩下的样本作为测试集. 相应的, 视角交叉主要在视角 2 和 3 中拍摄的样本进行模型训练, 在第 1 个视角样本进行测试.

除了以上列举的公共数据库外, 还有其他的一些比较有意义的 RGB-D 行为数据库, 本文仅在表 1 中给出一些简要的对比信息, 有兴趣的研究者可以到相关论文中了解更多详情.

2 RGB-D 行为识别模型介绍

由 Kinect 拍摄的 RGB-D 行为数据与传统的 RGB 视频数据具有很大的不同, 其主要包含深度视频、三维骨架和彩色视频三种模态的数据, 每种数据有很大的不同, 从不同角度刻画了行为内容信息. 由于 RGB-D 视频数据中的彩色图像信息分辨率比较低, 导致单纯依靠 RGB 视频中的常用行为识别方法并不能得到比较理想的结果[26]. 因此, 现有的 RGB-D 行为识别系统需要针对 RGB-D 数据特点发展对应的模型方法. 接下来, 本文将按模型所使用的数据模态进行划分, 分别介绍 RGB-D 行为识别方法.

2.1 基于深度模态数据的行为识别模型

为了构建基于深度视频数据的行为识别模型, 一个很直接的方式就是将 RGB 图像视频中常用的特征描述方式拓展应用到深度图像视频中, 使得拓展后的特征描述能够比较好地描述图像中的几何形状信息. 这方面最具有代表性的工作是文献 [18, 20, 26−27]. 这些方法试图将图像中的 HOG (Histogram of oriented gradient) 特征拓展成 4 维空间中的带方向直方图特征 HON4D (Histogram of oriented normal), 该特征主要刻画场景中的曲面法向量在 4 维空间上的分布信息. 具体地, 该方法把深度图像内容看成一个三维空间上的曲面, 相应的深度视频则可以定义为随时间变化的曲面流 $z = f(x, y, t)$. 曲面流在 x, y, z, t 4 维空间上的法向量为 $\boldsymbol{n} = (\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial z}{\partial t}, -1) = (f_x, f_y, f_z, -1)$. 法向量单位化之后, 可以表示为 $\boldsymbol{n} = (f_x, f_y, f_z, -1)/(f_x^2 + f_y^2 + f_z^2 + 1)$. 通过统计深度视频中 4 维法向量 $\boldsymbol{n}$ 在每个投影区间上的频率信息, 可以得到视频的直方图特征. 为了得到更加具有判别性的直方图特征, 作者们同时提出了一种可学习的直方图编码方式以自适应地确定投影区间. Liu 等[27] 对该方法进行了进一步的拓展, 通过计算局部深度时空方体中的法向量直方图以得到更多的有效几何信息.

Wang 等在文献 [2] 中通过随机采样大量的局部深度视频方体, 计算每个方体内包含点云[1] 的个数来刻画场景下的点云几何分布. Lu 等[28] 则直接比较随机采样得到的像素对之间的深度大小关系来表示形状. 这些方法试图通过对场景中的几何形状信息进行建模, 获取行为实施过程中的动作变化信息. 在建模过程中, 这些方法都忽略了纹理和人体姿势信息, 加上 Kinect 获取的深度数据具有比较多的噪声, 从而导致这些模型在很多数据库上的效果并不是特别理想.

2.2 基于三维骨架模态数据的行为识别模型

得益于微软开发的三维骨架实时捕捉系统[29], 系统可以比较准确地从深度视频数据中获取场景中行为人的三维骨架信息. 研究发现, 动态的三维骨架序列数据也能比较好地用来表示人体动作信息. 而且, 在建模过程中, 基于三维骨架构建的行为识别模型具有一定的鲁棒性, 不受纹理、背景等可能与行为无关因素的影响. 该类方法主要致力于挖掘各个关键骨架点位置[19, 30−32], 或者骨架点之间相对位置[33−35], 或者它们的组合[22, 36−37] 的动态信息进行识别. 在深度学习应用于 RGB-D 行为识别之前, 傅里叶变换被广泛用来提取骨架序列的动态信息[1, 12−13], 在建模过程中, 每个特征维度随着时间变化信息被当成一个单独的时间序列分别提取对应的傅里叶低频信息. 文献 [38] 通过将身体部位的位置信息投射到高维李群空间, 运用李代数中的运算技巧从时间序列数据中挖掘动态信息. 上述基于手工设计特征的算法往往不能捕捉到判别性的动作信息, 因而在很多行为数据库中的效果不是很理想.

近几年随着 GPU 计算能力的提升, 以及大规模 RGB-D 行为数据库 (NTU 大规模数据库[25]) 的出现, 涌现出了大量基于深度学习的方法, 应用最广泛的是循环神经网络 (Recurrent neural network, RNN) 和卷积神经网络 (Convolution neural network, CNN). LSTM (Long short-term memory)

[1] 即将深度图像像素点以三维坐标的形式展示.

作为 RNN 的一种变体在处理长时间序列数据时有着得天独厚的优势, 其能捕捉序列在较长时间内的相关性. Du 等[39] 在 2015 就使用 LSTM 建立编码器对骨架动作进行识别和预测. 其计算公式如下:

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{i}_t\\ \boldsymbol{f}_t\\ \boldsymbol{o}_t\\ \boldsymbol{u}_t\end{pmatrix}=\begin{pmatrix}\sigma\\ \sigma\\ \sigma\\ \tanh\end{pmatrix}\left(M\begin{pmatrix}\boldsymbol{x}_t\\ \boldsymbol{h}_{t-1}\end{pmatrix}\right)\\ &\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{u}_t\\ &\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t)\end{aligned}\tag{1}$$

其中, $\boldsymbol{x}_t$ 是 t 时刻的输入, $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{o}_t$, $\boldsymbol{u}_t$ 分别代表输入门、遗忘门、输出门和输入控制门, $\boldsymbol{c}_t$ 为细胞状态, 用来储存长期的时序信息, $\boldsymbol{h}_t$ 为隐藏状态, $\odot$ 代表元素乘积. LSTM 使用门来控制长时间信息和短时间信息的流动, 一定程度上消除了 RNN 的梯度消失问题, 所以能处理长时间的依赖关系.

然而传统的 RNN (LSTM) 忽视了骨架数据中的空间信息, 即骨架点间的相对位置. 文献 [32] 考虑人体骨架的空间结构, 将数据分为躯干和四肢 5 个部分, 分别使用 5 个双向循环神经网络 (Bidirectional recurrent neural network, BRNN) 提取特征, 然后将特征逐层合并送往下一层 BRNN 进行训练, 经过 4 层 BRNN 之后便完成了对人体各部位的空间关系从局部到整体的建模, 最后将整体的特征送入分类器进行分类, 具体网络结构如图 2. 这种分层 RNN 模型一定程度上挖掘了骨架数据的空间特征, 缺点在于由于模型过大, 参数量过多, 只有最后一层使用了双向 LSTM, 前面仅使用了普通的双向 RNN, 大大降低了模型的性能.

为了更加充分地挖掘骨架数据的空间信息, 文献 [40] 提出了时空 LSTM 模型. 传统的 LSTM 仅考虑时间维度使用细胞状态来储存长期的信息, 对于任意时刻的输入, 使用遗忘门、输入门和输出门丢弃或者增加信息, 而在时空 LSTM 中, 如图 3 所示, 当前时刻当前骨架点的状态 $\boldsymbol{h}_{j,t}$ 不仅与前一时刻的状态 $\boldsymbol{h}_{j,t-1}$ 有关, 还与前一骨架点的状态 $\boldsymbol{h}_{j-1,t}$ 有关, 作者使用两个遗忘门 $\boldsymbol{f}_{j,t}^S$ 和 $\boldsymbol{f}_{j,t}^T$ 分别控制时间和空间对当前状态的影响, 以此来同时挖掘空间特征和时间特征. 此外, 骨架点也不仅仅是按照传统顺序排列, 考虑到人体动作往往是由部分相邻的骨架点所决定的, 作者提出了循环遍历树结构进一步挖掘骨架点的空间信息. 时空 LSTM 计算流程如下:

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{i}_{j,t}\\ \boldsymbol{f}_{j,t}^S\\ \boldsymbol{f}_{j,t}^T\\ \boldsymbol{o}_{j,t}\\ \boldsymbol{u}_{j,t}\end{pmatrix}=\begin{pmatrix}\sigma\\ \sigma\\ \sigma\\ \tanh\end{pmatrix}\left(M\begin{pmatrix}\boldsymbol{x}_{j,t}\\ \boldsymbol{h}_{j-1,t}\\ \boldsymbol{h}_{j,t-1}\end{pmatrix}\right)\\ &\boldsymbol{c}_{j,t}=\boldsymbol{i}_{j,t}\odot\boldsymbol{u}_{j,t}+\boldsymbol{f}_{j,t}^S\odot\boldsymbol{c}_{t-1,t}+\boldsymbol{f}_{j,t}^T\odot\boldsymbol{c}_{j,t-1}\\ &\boldsymbol{h}_{j,t}=\boldsymbol{o}_{j,t}\odot\tanh(\boldsymbol{c}_{j,t})\end{aligned}\tag{2}$$

在行为识别中, 不同时刻的不同骨架点对于识别提供的信息量是非等同的, 所以注意力模型也被广泛地应用于此. 文献 [41] 分别使用两个网络来训练空域注意力模型和时域注意力模型, 空域注意力模型作用在网络的输入骨架点上, 时域注意力模型作用于主网络的输出特征上, 从而对不同时序和不同骨架点的信息进行加权, 最后实现端到端的行为识别. 可视化结果表明不同时域注意力模型会对更具判别力的帧赋予更大的权重, 对动作相关性较大的骨架点也会赋予更大的权重, 整体和人的感知一

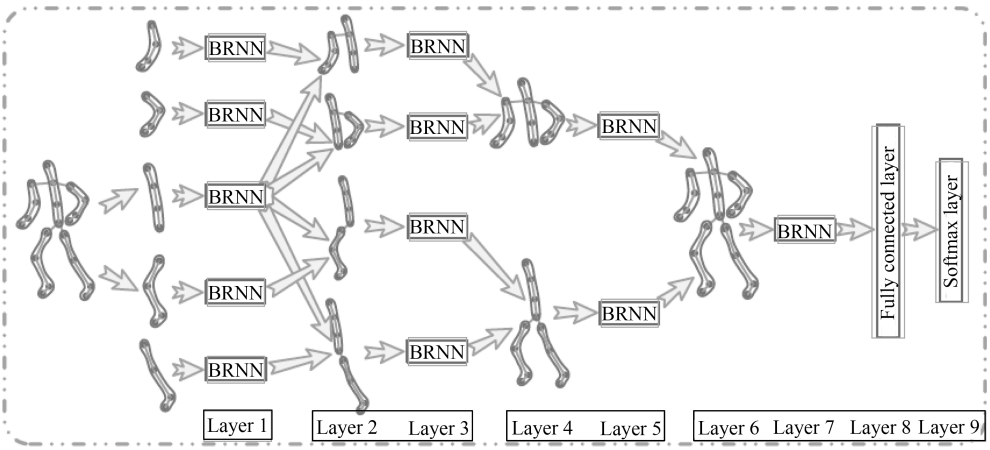


图 2　基于分层循环神经网络的三维骨架行为识别系统[32]

Fig. 2　Hierarchical recurrent neural network for skeleton based action recognition[32]

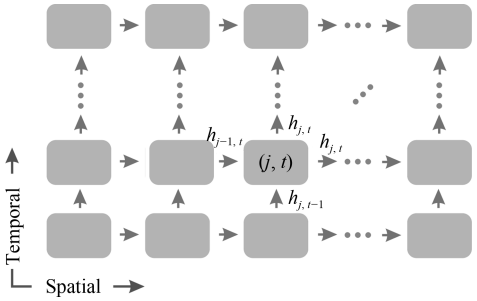


图 3 时空 LSTM[40]

Fig. 3 Spatio-temporal LSTM[40]

致, 图 4 为文献 [41] 在 "拳击" 这一动作中, 不同时刻不同节点的数据对目标行为的重要程度.

此外文献 [42] 还使用 LSTM 训练三维空间下的坐标变换矩阵, 以此来获取最佳坐标系下的骨架数据, 进而提升识别性能, 如图 5 所示三维欧氏空间下的坐标变换可以使用一个旋转矩阵 R_t 和一个平移向量 d_t 表示. 绕 Z 轴旋转 β 弧度的坐标变换矩阵为:

$$R_\beta^z = \begin{bmatrix} \cos(\beta) & \sin(\beta) & 0 \\ -\sin(\beta) & \cos(\beta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

以上均为使用循环神经网络 RNN 对时序数据进行建模的方法, 由于骨架数据不仅存在时间维度, 也存在空间维度, 所以一个骨架数据可以使用一个二维的矩阵来储存. 而近几年 CNN 模型在图像识别, 目标检测等领域中愈发成熟, 所以近两年也出现了很多基于 CNN 的特征提取模型, 取得了甚至比 RNN 更好的识别效果. 例如文献 [43] 首先计算全部骨架点与 4 个重要骨架点的相对距离, 然后将三维的笛卡尔坐标转化为球坐标, 经过双线性插值得到若干个固定大小的图片, 使用在 ImageNet 上预训练好的 VGG19[44] 模型提取特征, 经过时域中值池化后再使用全连接层 (Fully connected layer, FC) 进行分类, 在多个库中均取得了比基于 RNN 的识别算法更好的效果.

不同的行为中与之密切相关的骨架点也有所不同, 所蕴含的局部特征也不同, 这种特征称之为共现特征. 传统的 CNN 使用卷积核挖掘局部信息, 但只有卷积核内的相邻骨架点才被认为是在学习共现特征, 文献 [45] 提出了一种端到端的共现特征学习框架, 它首先在时间尺度上学习每个骨架点的特征, 然后将输出进行转置, 将骨架点维度和通道维度互换, 在后续的卷积层中聚合了所有关节的全局特征, 实验表明这种方法能比传统的 CNN 挖掘更多的共现信息.

总之, 数据驱动的深度学习方法给骨架行为识别领域带来了长足的进步, 虽然骨架数据并非传统意义上的图像, 但 CNN 强大的特征提取能力也使其越来越受研究者青睐. 此外, 在迁移学习的帮助下, 预训练好的神经网络比如 VGG、ResNet[46] 等能大大提升网络的训练速度, 相信深度学习方法在该研究领域还会有更进一步的突破.

2.3 基于多模态融合的行为识别模型

基于 RGB-D 视频融合的模型主要难点在于怎么去融合从不同模态数据中得到的特征. 当然不同模态特征的设计对融合系统的识别效果影响很大, 不同的融合系统对特征的要求也不一. 本节主要介绍基于多模态特征融合的识别方法.

文献 [12] 采用了从三维深度图像和骨架点提取到的两种特征: 深度局部占有信息和 3D 骨架点不变特征. 从深度图像提取局部占有特征的过程如下: 1) 针对每个骨架点, 从三维深度图像中提取其邻近的局部方体; 2) 按 x, y, z 轴方向分别将该局部方体分成 $N_x \times N_y \times N_z$ 个空间网格 (bins); 3) 针对每个网格 bin_{xyz}, 计算网格包含的像素点的个数并利用 Sigmoid 函数对其进行规则化, 最后得到每个 bin 的特征表示. 将所有 bin 的

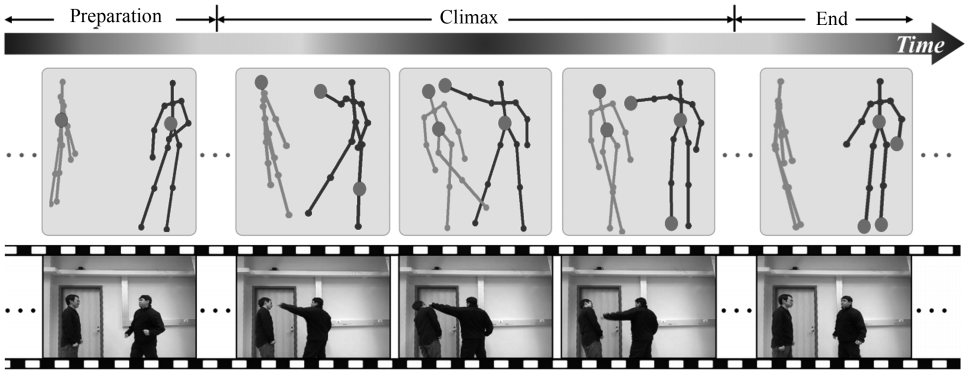


图 4 不同时刻不同节点和行为的相关程度[41]

Fig. 4 The correlation between different skeleton joints and actions at different moments[41]

特征表示串接到一起组成对当前帧的深度信息的局部占有特征; 4) 将视频序列的所有帧的局部占有特征当成一个时间序列, 提取其对应的傅里叶时域金字塔 (Temporal pyramid Fourier, TPF) 低频信息作为该节点的局部占有特征. 3D 节点不变特征的提取方法如下: 对于每个关节点, 首先计算它与其他节点的相对位置 (即 3 维坐标差), 然后提取其对应的金字塔傅里叶低频信息作为该节点的节点不变特征. 最后, 作者使用多核学习算法 (Multiple kernel learning, MKL) 挖掘出一些最具有代表性的骨架点进行融合实现对视频表示. 如图 6 所示, 该方法的优势在于其能结合深度特征和骨架特征, 利用判别学习方法从不同模态特征中选取出最有价值的行为特征. 然而, 它没有深入考虑不同特征之间的内在结构联系, 这个缺点限制了该方法在 RGB-D 行为识别方面的效果.

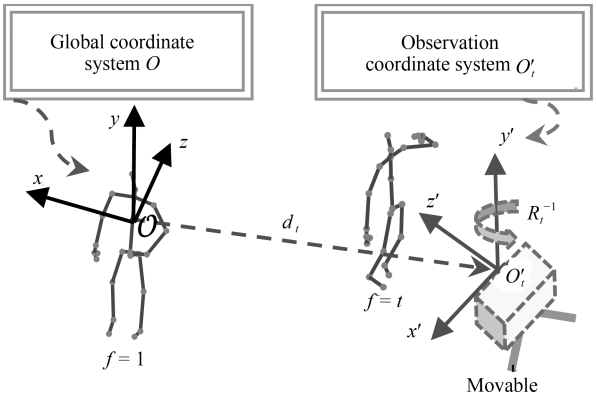


图 5　学习一个坐标转移矩阵转换骨架数据的坐标系[42]

Fig. 5　Learning an optimal coordinate transition matrix to transform the coordinate system[42]

文献 [47] 考虑挖掘多个身体部位的多模态特征之间内部的结构信息, 以选取到对识别最优的特征组合进行识别. 在建模过程中, 作者提出了一种层次混合范数, 对特征按部位和模态进行层次划分, 对不同的层次使用不同的规则范数进行归一化, 从而挖掘特征之间的结构联系.

文献 [1] 通过融合从 RGB、深度视频和 3 维骨架序列中提取到的动态特征实现识别. 其中 RGB、深度视频方面的动态特征构建如下: 1) 从骨架点对应的人体部位周围提取深度 (彩色) 图像 HOG 特征; 2) 提取视频对应的 HOG 特征序列的傅里叶低频系数作为特征表示. 考虑到不同模态的特征具有一定的异质性, 即特征具有不同的维度, 不同的性质. 作者通过提出一种基于多任务学习的异质特征学习模型来挖掘不同特征之间的共享成分和私有成分 (图 7), 在多个数据库中 (如 SYSU 3DHOI[1], MSRDaily[12] 和 CAD60[15]) 达到比较好的识别效果. 同时, 作者还发现, 通过引入迁移学习技术, 利用其他行为数据库作为辅助库可以稳定提升目标数据库中的特征学习效果.

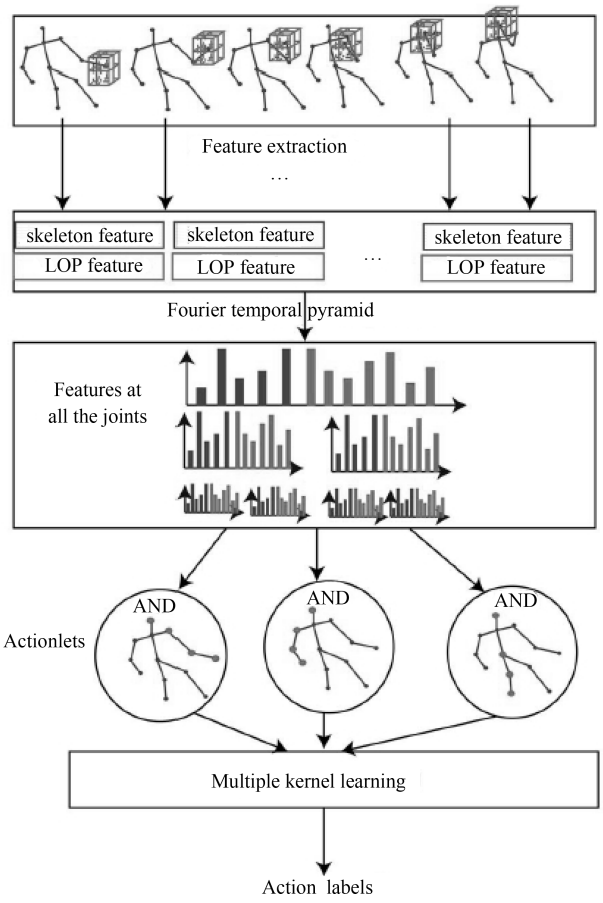


图 6　学习判别 Actionlet 集合进行行为识别[1]

Fig. 6　Learning actionlet ensemble for 3D human action recognition[1]

在文献 [1] 的基础上, 文献 [48] 进一步发展了一个共享–私有特征学习的深度学习框架. 在该深度框架里, 作者定义了一个网络层将模态特征分解为共享成分和私有成分. 为了能够提升学习效率, 作者对分解后的特征加以了稀疏的约束. 多个层次的组合构成了一个深度的共享–私有特征学习框架.

总体而言, 以上方法能较好地利用不同模态的特征数据, 且在现有 RGB-D 行为数据库上也取得了非常不错的结果. 大规模行为数据库的出现也促进了相关的深度学习算法发展, 但以数据和任务为驱动的深度学习技术并没有得到很好的应用, 现有工作基本把特征融合和特征提取分成了两个隔离的部分, 相互之间不能促进. 基于端到端的多模态特征融合技术是未来需要进一步发展的技术, 相信其在 RGB-D 行为识别中能够取得更好的结果.

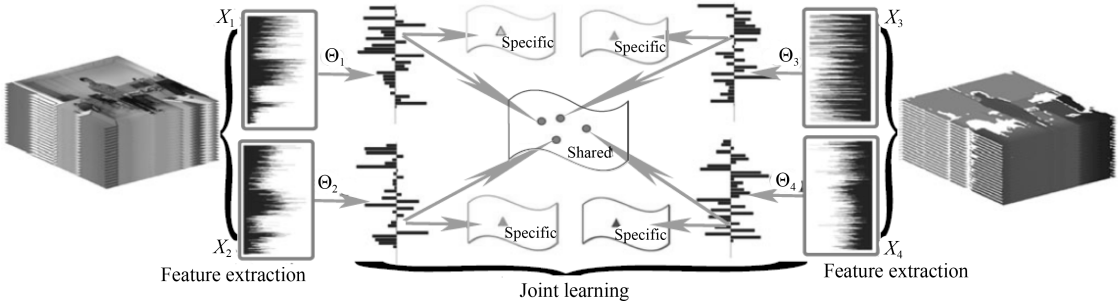


图 7 多模态异质特征共享结构与私有结构同步学习模型[1]

Fig. 7 Jointly learning heterogeneous features for RGB-D activity recognition[1]

在实际应用中可能出现部分模态数据丢失或者很难获取的情况. 针对部分模态数据缺失的多模态融合学习方法研究有着重要的意义. 例如, 在文献 [49] 中, 作者在模型训练的过程中引入了姿势 (骨架) 信息以学习到更合适的视频行为注意力 (Attention) 参数, 而在测试过程中不需要输入姿势信息.

3 现有 RGB-D 行为识别方法的实验对比与分析

前面章节主要介绍了 RGB-D 行为识别领域中常用的公共数据库和近些年来提出的相关识别方法及其发展. 本节将结合 NTU 数据库, MSR 日常行为数据库和 SYSU 3DHOI 数据库具体对比分析相关识别模型.

表 2～表 4 分别给出了相关方法在 NTU 大规模行为数据库, MSR 日常行为数据库和 SYSU 3DHOI 数据库上的识别结果. 从中可以看出, 自深度学习方法被广泛用于解决 RGB-D 行为识别问题以来, 具体识别效果有了大幅度的提高, 尤其是在 NTU 大规模行为数据库上, 无论是个体交叉还是视角交叉设置, 现有方法仅使用骨架数据就能将别性能从 60 % 提升至 90 % 左右. 其中大部分的深度学习相关工作都是基于改进 LSTM 模型, 以挖掘动作序列中的时空变化信息. 虽然 LSTM 模型充分展现了它在时序建模方面的强大能力, 但不能忽视的是, 最新的一些研究表明基于卷积神经网络 (CNN) 的模型也取得了非常优异的识别结果[43, 45], 通过将三维骨架序列人工编码成静态图像, 利用卷积核自动学习图像内部编码的时空结构信息, 从而挖掘到具有判别性的时空变化信息. 然而值得注意的是, 这些模型需要人工地将三维骨架序列进行编码, 且实验表明该编码方式对算法的识别效果较大. 因此, 怎样对三维骨架序列进行合适编码, 是该研究中的关键问题. 另一方面, 从表 3 和表 4 的识别结果可以看到, 基于 RGB-D 的多数据模态融合模型往往比单一模态方法识别效果更加稳定. 这很符合预期, 因为不同模态数据可以捕捉到行为不同方面的信息, 它们之间往往能在一定程度上进行互补. 然而, 由于从多个通道提取特征非常消耗计算资源和耗时, 尤其是当使用深度学习网络提取相关特征时. 这也导致大部分的多模态特征融合方法在 NTU 大规模数据库上未能进行验证. 因此, 怎样发展一个轻

表 2 在 NTU RGB-D 数据库上各种方法的识别结果对比 ("RGB-D" 指同时使用 RGB、深度和骨架三种模态数据)

Table 2 Comparison of action recognition accuracies on the NTU RGB-D dataset ("RGB-D" indicates that the approach employs all the RGB, depth, and skeleton modalities for recognition)

方法	数据模态	准确度 (%)	
		个体交叉	视角交叉
HON4D[29]	深度	30.6	7.3
Skeletal quads[50]	骨架	38.6	41.4
Lie group[37]	骨架	50.1	52.8
Hierarchical RNN[39]	骨架	59.1	64.0
Deep RNN[17]	骨架	59.3	64.1
Dynamic skeletons[13]	骨架	60.2	65.2
Deep LSTM[17]	骨架	60.7	67.3
Part-aware LSTM[17]	骨架	62.9	70.3
ST-LSTM[40]	骨架	65.2	76.1
ST-LSTM + Trust gate[40]	骨架	69.2	77.7
STA-LSTM[41]	骨架	73.4	81.2
Deep multimodal[48]	RGB-D	74.9	—
Multiple stream[51]	RGB-D	79.7	81.43
Skeleton and depth[52]	深度 + 骨架	75.2	83.1
Clips+CNN+MTLN[43]	骨架	79.6	84.8
VA-LSTM[42]	骨架	79.4	87.6
Pose-attention[53]	RGB + 骨架	82.5	88.6
Deep bilinear[54]	RGB-D	85.4	90.7
HCN[45]	骨架	86.5	91.1
DA-Net[55]	RGB	88.12	91.96
SR-TSL[56]	骨架	84.8	92.4

表 3 在 MSR 数据库上各种方法的识别结果对比

Table 3 Comparison of action recognition accuracies on the MSR daily activity dataset

方法	数据模态	准确度 (%)
Dynamic temporal warping[57]	骨架	54
3D Joints and LOP Fourier[12]	深度 + 骨架	78
HON4D[23]	深度	80.00
SSFF[58]	RGB-D	81.9
HFM[1, 59]	RGB-D	84.38
Deep model (RGGP)[60]	RGB-D	85.6
Actionlet ensemble[12]	深度 + 骨架	85.75
Super normal[19]	深度	86.25
Bilinear[61]	深度	86.88
DCSF + Joint[62]	RGB-D	88.2
MPCCA[1, 63]	RGB-D	90.62
MTDA[1, 64]	RGB-D	90.62
LFF + IFV[65]	骨架	91.1
Group sparsity[33]	骨架	95
JOULE[1]	RGB-D	95
Range sample[28]	深度	95.6
Deep multi modal[48]	RGB-D	97.5

表 4 在 SYSU 3D HOI 数据库上各种方法的识别结果对比 ("RGB-D" 指同时使用 RGB、深度和骨架三种模态数据)

Table 4 Comparison of action recognition accuracies on the SYSU 3D HOI Dataset ("RGB-D" indicates that the approach employs all the RGB, depth, and skeleton modalities for recognition)

方法	数据模态	准确度 (%)	
		设置 1	设置 2
HON4D[1, 23]	深度	73.4	79.2
HFM[1, 59]	RGB-D	75	76.7
ST-LSTM[40]	骨架	76.5	—
VA-LSTM[42]	骨架	76.9	77.5
MPCCA[1, 63]	RGB-D	76.3	80.7
SR-TSL[56]	骨架	80.7	81.9
MTDA[1, 64]	RGB-D	79.2	84.2
JOULE[1]	RGB-D	79.6	84.9

量级的深度学习模型来融合 RGB、深度和三维骨架数据进行行为识别也是未来的一个重要研究内容.

4 思考与展望

基于 RGB-D 的人体行为识别一直是计算机视觉领域的热点问题, 近几年随着深度学习的兴起, RGB-D 行为识别领域有了很大的突破, 通过神经网络技术以数据驱动方式自动学习到的特征逐渐代替了 HOG, SIFT 等手工设计特征, 相关大规模行为数据集的出现进一步推动了基于深度学习的识别算法的发展. 特别地, 卷积神经网络 (CNN) 的 RGB-D 行为识别模型在部分行为数据库上已经达到了相当高的识别率. 然而仍存在着不少问题有待解决. 首先, 在 RGB-D 行为识别中, 深度视频、RGB 视频以及骨架三种特征提取和网络训练都需要耗费大量的时间和计算资源, 如何高效地进行多模态特征融合就显的尤为重要. 尽管文献 [52, 54] 通过利用双线性池化[66] 操作一定程度上提升了融合效率, 但仍有很大的提升空间, 多模态行为识别仍有待进一步研究. 其次, 实际测试中往往可能会遇到部分模态数据缺失或失效的情况, 怎么调整多模态融合学习算法使得其能充分利用获取到的部分模态数据, 也是一个重要的需要解决的研究内容. 最后, 在数据库设计方面, 现有的 RGB-D 行为数据库都主要记录室内控制场景下的人体行为, 行为样本缺少多样性, 期待未来有更加复杂的大规模 RGB-D 行为数据库的出现.

考虑到行为识别是 "事后" 识别研究, 即系统需要在行为动作完成之后再进行识别. 面向正在进行的部分行为的 RGB-D 行为前期预测问题也逐渐受到了众多研究者的关注[67−68]. 在无人驾驶、机器人以及医疗监控等很多应用场景下, 人们更希望在动作实施完成前系统便能及时地预测和识别, 这可以给我们足够的反应时间来提前做好准备. 例如, 当系统观测到一个患者失去了平衡时, 可能即将会摔倒, 我们希望系统能及时预测到这一动作的发生, 并做出相应的反应. 早期的工作主要基于马尔科夫模型 (Markov model, MM)、条件随机场 (Conditional random fields, CRF) 等非深度学习方法, 近几年则主要是利用 RNN 和生成对抗网络 (Generative adversarial networks, GAN). 文献 [68] 针对不完整的视频学习一个弱类标, 从而可以利用部分视频和完整视频中学习到一个鲁棒的 RNN 行为预测器. 文献 [38] 利用 RNN 配合编码器和解码器, 通过最小化预测值和真实值之间的欧氏距离来训练网络. 文献 [69] 采用 GAN 模型, 通过同时训练生成器和判别器来预测骨架的行为特征. 文献 [70] 提出时域自适应选择网络, 同步学习行为起始时间和行为预测器, 从而实现从未切割的长视频中预测行为类别. 总体而言, 基于 RGB-D 视频数据的前期行为预测无论在研究还是应用方面, 未来都有很大的发展空间.

5 结论

本文详细介绍了 RGB-D 行为识别领域中具有代表性的数据库, 然后根据使用的数据模态类型对现有研究方法进行划分, 分别介绍了基于深度数据、基于三维骨架数据以及基于多模态融合的 RGB-D 行为识别研究进展. 基于传统机器学习方法的识别算法采用手工设计特征挖掘人体的运动信息并进行分类, 在数据库较小时能取得较好的效果, 但在面对复杂的数据库时分类效果就未尽人意. 而数据驱动的深度学习方法需要大量已知标签的数据进行训练, 可以自发地学习人体行为特征, 所以在复杂问题

面前有着比传统方法更好的效果. 但同时我们也应该注意到深度学习方法需要大量的数据和较长的训练时间, 基于深度学习和多模态特征融合的 RGB-D 行为识别方法在计算效率上也有待进一步提升.

References

1 Hu J F, Zheng W S, Lai J H, Zhang J G. Jointly learning heterogeneous features for RGB-D activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(11): 2186−2200

2 Wang J, Liu Z C, Chorowski J, Chen Z Y, Wu Y. Robust 3D action recognition with random occupancy patterns. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 872−885

3 Liu Zhi, Dong Shi-Du. Study of human action recognition by using skeleton motion information in depth video. *Computer Applications and Software*, 2017, **34**(2): 189−192, 219 (刘智, 董世都. 利用深度视频中的关节运动信息研究人体行为识别. 计算机应用与软件, 2017, **34**(2): 189−192, 219)

4 Wang Song-Tao, Zhou Zhen, Qu Han-Bing, Li Bin. Bayesian saliency detection for RGB-D images. *Acta Automatica Sinica*, 2017, **43**(10): 1810−1828 (王松涛, 周真, 曲寒冰, 李彬. RGB-D 图像的贝叶斯显著性检测. 自动化学报, 2017, **43**(10): 1810−1828)

5 Wang Xin, Wo Bo-Hai, Guan Qiu. Human action recognition based on manifold learning. *Chinese Journal of Image and Graphics*, 2014, **19**(6): 914−923 (王鑫, 沃波海, 管秋. 基于流形学习的人体动作识别. 中国图象图形学报, 2014, **19**(6): 914−923)

6 Liu Xin, Xu Hua-Rong, Hu Zhan-Yi. GPU based fast 3D-object modeling with Kinect. *Acta Automatica Sinica*, 2012, **38**(8): 1288−1297 (刘鑫, 许华荣, 胡占义. 基于 GPU 和 Kinect 的快速物体重建. 自动化学报, 2012, **38**(8): 1288−1297)

7 Wang Liang, Hu Wei-Ming, Tan Tie-Niu. A survey of visual analysis of human motion. *Chinese Journal of Computers*, 2002, **25**(3): 225−237 (王亮, 胡卫明, 谭铁牛. 人运动的视觉分析综述. 计算机学报, 2002, **25**(3): 225−237)

8 Georgia Gkioxari, Ross Girshick, Piotr Dollár, Kaiming He. Detecting and Recognizing Human-Object Interactions. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018. DOI: 10.1109/CVPR.2018.00872

9 Kläser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the 2008 British Machine Vision Conference. Leeds, UK: British Machine Vision Association, 2008.

10 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

11 Wang X Y, Han T X, Yan S C. An HOG-LBP human detector with partial occlusion handling. In: Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 32−39

12 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

13 Hu J F, Zheng W S, Lai J H, Zhang J G. Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 5344−5352

14 Wei P, Zhao Y B, Zheng N N, Zhu S C. Modeling 4D human-object interactions for event and object recognition. In: Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Sydney: IEEE, 2013. 3272−3279

15 Sung J, Ponce C, Selman B, Saxena A. Human activity detection from RGBD images. In: Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition. San Francisco, USA: AAAI, 2011. 47−55

16 Koppula H S, Gupta R, Saxena A. Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, 2013, **32**(8): 951−970

17 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, USA: IEEE, 2016.

18 Zhu Y, Chen W B, Guo G D. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing*, 2014, **32**(8): 453−464

19 Yang X D, Tian Y L. Super normal vector for activity recognition using depth sequences. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 804−811

20 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

21 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. San Francisco, USA: IEEE, 2010. 9−14

22 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. Providence, USA: IEEE, 2012. 20−27

23 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. Portland, USA: IEEE, 2013. 716−723

24 Ni B B, Wang G, Moulin P. RGBD-HuDaAct: a color-depth video database for human daily activity recognition. *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. London, UK: Springer, 2013. 193−208

25 Lillo I, Soto A, Niebles J C. Discriminative hierarchical modeling of spatio-temporally composable human activities. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 812−819

26 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. Singapore: Springer, 2014. 50−65

27 Liu A A, Nie W Z, Su Y T, Ma L, Hao T, Yang Z X. Coupled hidden conditional random fields for RGB-D human action recognition. *Signal Processing*, 2015, **112**: 74−82

28 Lu C W, Jia J Y, Tang C K. Range-sample depth feature for action recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 772−779

29 Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, et al. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013, **56**(1): 116−124

30 Hussein M E, Torki M, Gowayyed M A, El-Saban M. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: AAAI, 2013. 2466−2472

31 Lv F J, Nevatia R. Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost. In: Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer, 2006. 359−372

32 Yang X D, Tian Y L. EigenJoints-based action recognition using naive-Bayes-nearest-neighbor. In: Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 14−19

33 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. Sydney, Australia: IEEE, 2013. 1809−1816

34 Ofli F, Chaudhry R, Kurillo G, et al. Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 2014, **25**(1): 24−38

35 Zhu Y, Chen W B, Guo G D. Fusing spatiotemporal features and joints for 3D action recognition. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, USA: IEEE, 2013. 486−491

36 Zanfir M, Leordeanu M, Sminchisescu C. The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 2752−2759

37 Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a Lie group. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 588−595

38 Fragkiadaki K, Levine S, Felsen P, Malik J. Recurrent network models for human dynamics. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4346−4354

39 Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1110−1118

40 Liu J, Shahroudy A, Xu D, Kot A C, Wang G. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(12): 3007−3021

41 Song S J, Lan C L, Xing J L, Zeng W J, Liu J Y. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017. 4263−4270

42 Zhang P F, Lan C L, Xing J L, Zeng W J, Xue J R, Zheng N N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2136−2145

43 Ke Q H, Bennamoun M, An S J, Sohel F, Boussaid F. A new representation of skeleton sequences for 3D action recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4570−4579

44 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.

45 Li C, Zhong Q Y, Xie D, Pu S L. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv preprint arXiv:1804.06055, 2018.

46 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

47 Shahroudy A, Ng T T, Yang Q X, Wang G. Multimodal multipart learning for action recognition in depth videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(10): 2123−2129

48 Shahroudy A, Ng T T, Gong Y H, Wang G. Deep multimodal feature analysis for action recognition in RGB+D videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(5): 1045−1058

49 Du W B, Wang Y L, Qiao Y. RPAN: an end-to-end recurrent pose-attention network for action recognition in videos. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 3745−3754

50 Evangelidis G, Singh G, Horaud R. Skeletal quads: human action recognition using joint quadruples. In: Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014. 4513−4518

51 Garcia N C, Morerio P, Murino V. Modality distillation with multiple stream networks for action recognition. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018.

52 Rahmani H, Bennamoun M. Learning action recognition model from depth and skeleton videos. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 5833−5842

53 Baradel F, Wolf C, Mille J. Human action recognition: pose-based attention draws focus to hands. In: Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy: IEEE, 2017.

54 Hu J F, Zheng W S, Pan J H, Lai J H, Zhang J G. Deep bilinear learning for RGB-D action recognition. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018.

55 Wang D A, Ouyang W L, Li W, Xu D. Dividing and aggregating network for multi-view action recognition. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018.

56 Si C Y, Jing Y, Wang W, Wang L, Tan T N. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018.

57 Müller M, Röder T. Motion templates for automatic classification and retrieval of motion capture data. In: Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Vienna, Austria: Eurographics Association Aire-la-Ville, 2006. 137−146

58 Shahroudy A, Wang G, Ng T T. Multi-modal feature fusion for action recognition in RGB-D sequences. In: Proceedings of the 6th International Symposium on Communications, Control and Signal Processing (ISCCSP). Athens, Greece: IEEE, 2014. 1−4

59 Cao L L, Luo J B, Liang F, Huang T S. Heterogeneous feature machines for visual recognition. In: Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 1095−1102

60 Liu L, Shao L. Learning discriminative representations from RGB-D video data. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: AAAI, 2013. 1493−1500

61 Kong Y, Fu Y. Bilinear heterogeneous information machine for RGB-D action recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1054−1062

62 Xia L, Aggarwal J K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013. 2834−2841

63 Cai Z W, Wang L M, Peng X J, Qiao Y. Multi-view super vector for action recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 596−603

64 Zhang Y, Yeung D Y. Multi-task learning in heterogeneous feature spaces. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2011.

65 Yu M Y, Liu L, Shao L. Structure-preserving binary representations for RGB-D action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(8): 1651−1664

66 Gao Y, Beijbom O, Zhang N, Darrell T. Compact bilinear pooling. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 317−326

67 Hu J F, Zheng W S, Ma L Y, Gang W, Lai J H, Zhang J G. Early action prediction by soft regression. In: Proceedings of the 2018 IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 1−1

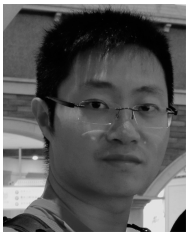
68 Hu J F, Zheng W S, Ma L Y, Wang G, Lai J H. Real-time RGB-D activity prediction by soft regression. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 280−296

69 Barsoum E, Kender J, Liu Z C. HP-GAN: probabilistic 3D human motion prediction via GAN. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, USA: IEEE, 2018.

70 Liu J, Shahroudy A, Wang G, Duan L Y, Kot A C. SSNet: scale selection network for online 3D action prediction. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018.

胡建芳 中山大学副研究员. 2016 年获得中山大学数学系博士学位. 主要研究方向为计算机视觉与模式识别.
E-mail: hujf5@mail.sysu.edu.cn
(**HU Jian-Fang** Associate professor at Sun Yat-sen University. He received his Ph. D. degree from the School of Mathematics, Sun Yat-Sen University in 2016. His research interest covers computer vision and patten recognition.)

王熊辉 中山大学模式识别与智能系统专业硕士研究生. 2015 年获得中山大学智能科学与技术学士学位. 主要研究方向为图像处理, 计算机视觉与模式识别.
E-mail: wxiongh@mail2.sysu.edu.cn
(**WANG Xiong-Hui** Master student at Sun Yat-sen University. He received his bachelor degree in intelligence science and technology from Sun Yat-sen University in 2015. His research interest covers image processing, computer vision, and pattern recognition.)



郑伟诗 中山大学数据科学与计算机学院教授. 他已发表 100 余篇主要学术论文, 其中 70 余篇发表在图像识别和模式分类领域 IEEE TPAMI, IEEE TIP, IEEE TNNLS 等国际主流权威期刊和 ICCV, CVPR 等计算机学会推荐 A 类国际学术会议. 担任 Pattern Recognition 等期刊的编委, 担任 AVSS2012, ICPR2018, BMVC2018 Area Chair 等. 主要研究方向为视频监控下的行人身份识别与行为信息理解.
E-mail: zhwshi@mail.sysu.edu.cn
(**ZHENG Wei-Shi** Professor at Sun Yat-sen University. His research interest covers person re-identification, action/activity recognition. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme in 2015. He is an associate editor of Pattern Recognition. He is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of Royal Society-Newton Advanced Fellowship, United Kingdom.)

赖剑煌 中山大学教授. 1999 年获得中山大学数学系博士学位. 目前在 IEEE TPAMI, IEEE TNNLS, IEEE TIP, IEEE TSMC-B, PR, ICCV, CVPR, and ICDM 等国际权威刊物发表论文 200 多篇. 主要研究方向为图像处理, 计算机视觉, 模式识别. 本文通信作者.
E-mail: stsljh@mail.sysu.edu.cn
(**LAI Jian-Huang** Professor at Sun Yat-Sen University. He received his Ph. D. in mathematics from Sun Yat-Sen University in 1999. He has published over 200 scientific papers in international journals and conferences including IEEE TPAMI, IEEE TNNLS, IEEE TIP, IEEE TSMC-B, PR, ICCV, CVPR, and ICDM. His research interest covers digital image processing, computer vision, pattern recognition. Corresponding author of this paper.)