计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):146−152 [doi: 10.15888/j.cnki.csa.007875]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多模态特征学习的人体行为识别方法①

周雪雪, 雷景生, 卓佳宁

(上海电力大学 计算机科学与技术学院, 上海 200090)
通讯作者: 周雪雪, E-mail: zhouxuexue5@163.com

摘　要: 由于从单一行为模态中获取的特征难以准确地表达复杂的人体动作, 本文提出基于多模态特征学习的人体行为识别算法. 首先采用两条通道分别提取行为视频的 RGB 特征和 3D 骨骼特征, 第 1 条通道 C3DP-LA 网络由两部分组成: (1) 包含时空金字塔池化 (Spatial Temporal Pyramid Pooling, STPP) 的改进 3D CNN; (2) 基于时空注意力机制的 LSTM, 第 2 条通道为时空图卷积网络 (ST-GCN), 然后, 本文将提取到的两种特征融合使其优势互补, 最后用 Softmax 分类器对融合特征进行分类, 并在公开数据集 UCF101 和 NTU RGB + D 上验证. 实验表明, 本文提出的方法与现有行为识别算法相比具有较高的识别准确度.

关键词: 行为识别; 改进 3D CNN; 时空注意力; 时空图卷积网络; 特征融合

引用格式: 周雪雪,雷景生,卓佳宁.基于多模态特征学习的人体行为识别方法.计算机系统应用,2021,30(4):146–152. http://www.c-s-a.org.cn/1003-3254/7875.html

Human Action Recognition Algorithm Based on Multi-Modal Features Learning

ZHOU Xue-Xue, LEI Jing-Sheng, ZHUO Jia-Ning

(College of Computer and Science, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: Since the features obtained from a single action mode fail to accurately express complex human actions, this study proposes a recognition algorithm for human actions based on multi-modal feature learning. First, two channels extract the RGB and 3D skeletal features from the action video. The first channel, i.e., the C3DP-LA network, consists of an improved 3D CNN with Spatial Temporal Pyramid Pooling (STPP) and LSTM based on spatial-temporal attention. The second channel is the Spatial-Temporal Graph Convolutional Network (ST-GCN). Then the two extracted features are fused and classified by Softmax. Furthermore, the proposed algorithm is verified on the public data sets UCF101 and NTU RGB+D. The results show that this algorithm has higher recognition accuracy than its counterparts.

Key words: action recognition; improved 3D CNN; Spatial-Temporal Attention (ST-Att); Spatial-Temporal Graph Convolutional Network (ST-GCN); feature fusion

近年来, 随着计算机视觉技术的不断发展, 人体行为识别逐渐成为一个重要的研究领域, 在视频监控、医疗看护、游戏应用与人机交互等方面有着广泛的应用[1]. 目前, 人类行为主要可以基于 RGB 视频[2,3], 深度图[4,5] 和 3D 骨架[6,7] 等 3 种模态的特征进行识别.

尽管基于每种特征的识别技术发展迅速并取得了很多成果, 当前仍然存在以下几个问题: (1) 现有的人体识别算法大多是基于单一模态特征进行识别的. (2) 基于 RGB 视频的行为识别容易受到遮挡、环境变化或阴影的干扰; 深度图中颜色和纹理的缺失容易导致相关模型识别率较低; 3D 骨架由于角度、姿势以及关节点数有限等原因, 容易导致动作被错检或漏检.

① 基金项目: 国家自然科学基金 (61672337)
Foundation item: National Natural Science Foundation of China (61672337)
收稿时间: 2020-08-25; 修改时间: 2020-09-15; 采用时间: 2020-09-25; csa 在线出版时间: 2021-03-30

(3) 视频中存在大量与行为识别无关的画面, 这些信息会降低算法的准确度. 针对以上情况, 本文融合 RGB 视频和 3D 骨架两种行为信息的特征, 充分利用两者的优势, 同时引用注意力机制来研究行为识别.

对于 RGB 视频行为特征, 前人提出了一些经典的识别模型. Tran 等[8] 采用 3D 卷积和 3D 池化构建了三维卷积神经网络 (3D CNN), 它可以同时提取视频行为的外观和运动特征, 而且结构简单, 运行速度较大多行为识别算法更快, 并在 UCF101 等公开数据集上取得了不错的效果. 然而, 3D CNN 也存在一定的技术缺陷: (1) 网络的训练及测试均要求输入尺寸和比例固定的视频帧. 当输入任意大小的视频时, 3D CNN 会对其进行裁剪或缩放以产生固定大小的输入样本, 而这种操作会导致信息丢失或扭曲, 从而影响特征的提取, 如图 1 所示. (2) 网络每次只能接收 7 帧输入. 3D CNN 将连续的视频分割成多个长度为 7 帧的片段, 降低了动作识别的连续性, 具有一定的局限.

(a) 裁剪 (b) 缩放

图 1 将视频帧裁剪或缩放后导致关键信息丢失

如今, He 等[9] 提出的空间金字塔池化网络 (SPP-net) 已经成功解决了深度神经网络中输入数据维度固定的问题, 并在目标分类、目标检测等领域取得了良好的效果. 本文将空间金字塔池化扩展为时空金字塔池化 (STPP), 并将其应用在 3D CNN 中, 使得任意尺寸的视频都可以直接输入网络, 并产生固定大小的输出. 此外, LSTM 因其对长短时特征的记忆功能而被广泛应用于视频识别中, 由于 3D CNN 不能充分提取长时序的时间特征, 本文采用添加时空注意力机制[10] 的 LSTM 来进一步获取长时序视频帧的时间信息, 并自适应地分配不同注意力的权重, 感知关键帧信息, 最终得到更为完整的动态行为.

基于骨架的行为识别方面, 本文采用 Yan 等[11] 提出的时空图卷积网络提取骨骼特征. 在骨骼序列上构建时空图, 通过对其应用多层时空图卷积操作, 逐渐在图像上生成高级的骨骼特征. 最后, 本文将第 1 层通道 C3DP-LA 提取到的 RGB 视频特征和第 2 层通道 ST-GCN 提取到的骨骼特征进行早期融合, 充分学习不同类型特征的优点, 并用标准的 Softmax 分类器完成动作识别.

本文的贡献: (1) 考虑到单一模态的特征各有不足, 本文提出一种双流行为识别框架, 先分别提取两种不同类型的特征, 再将其融合, 利用两者的互补性综合表征人体行为. (2) 为了能够处理任意大小和长度的 RGB 视频, 本文在 3D CNN 中接入时空金字塔池化, 然后连接 LSTM 学习时间特征. (3) 为了增强关键特征, 提高算法精度, 本文在 LSTM 模块加入时空注意力机制. (4) 本文的方法在 NTU RGB+D 数据集上优于现有的一些算法, 在基于单一特征和融合特征两类识别方法中表现出良好的识别效果.

1 相关工作

人体行为识别是计算机视觉领域中的一个热门研究课题. 目前, 针对动作识别的研究大多是基于单一模态开展的, 例如, Simonyan 等[12] 提出的首个双流卷积网络框架, 采用两个分支 CNN 分别对 RGB 视频的静态帧图像和动态光流进行特征提取, 以获得空间和时间信息, 最后用 SVM 将两种信息进行融合分类, 完成动作的识别. Chen 等[13] 提出基于深度图的行为识别算法 DMMs, 利用深度图投影之间的绝对差形成一个 DMM, 然后应用带有距离加权的正则协同分类器识别动作. Lee 等[14] 提出基于骨架进行动作识别的时间滑动 LSTM (TS-LSTM) 网络, 依靠多个 LSTM 的集合捕获人体行为的短期、中期和长期运动特性, 有效地学习时间和空间特征, 增加对动态时间变化的鲁棒性. 这些方法可以正确识别一些动作, 但单一模态的特征难以准确、全面地表达复杂的人体动作. 为了解决这一问题, 一些研究者尝试将不同模态的特征融合起来, 利用其互补性达到更好的识别效果.

Chaaraoui 等[15] 提出一种二维形状的人体姿态估计与骨骼特征相结合的方法, 通过将有效的 2D 轮廓和 3D 骨骼特征融合获取具有较高鉴别价值的视觉特征, 同时利用轮廓提供的额外判别数据, 提高人体行为识别误差的鲁棒性. Sanchez-Riera 等[16] 针对手势识别和通用对象识别, 将 RGB 特征与深度特征融合起来, 并评估早期和晚期融合两种方案, 结果表明, 两种特征的早期融合相比于晚期融合和单一特征具有更有效的行为表达能力. Li 等[17] 提出了多特征稀疏融合模型,

分别从骨架和深度数据中提取人体部位的多个特征, 并利用稀疏正则化技术自动识别关键部分的特征结构, 由此学习到的加权特征对于多任务分类更具鉴别性. Chen 等[18] 基于深度相机和惯性体传感器, 分别提取人体行为的深度图像特征和 RGB 视频特征, 并评估特征级融合和决策级融合两种识别框架.

上述多特征融合模型由于从所选模态中提取的时间或空间特征不够显著, 识别准确度仍然有所欠缺. 考虑到深度图像色彩、纹理等重要信息的缺失可能导致模型混淆分类, 本文从 RGB 视频和 3D 骨骼两种模态中提取特征, 将其融合, 利用两种特征的优势进行动作分类.

2 算法框架

本文基于多模态特征融合的行为识别算法框架如图 2 所示. 模型在行为特征的提取阶段主要包含两层: C3DP-LA 网络和 ST-GCN 网络, 其中 C3DP-LA 网络由改进 3D CNN 和包含时空注意力机制的 LSTM 模型组成. 输入一段视频, 我们可以得到行为样本的 RGB 视频序列和 3D 骨骼序列. 在 C3DP-LA 网络中, 添加时空金字塔池化的 3D CNN 可以自动处理任意尺寸大小的 RGB 视频帧序列, 快速卷积得到初步的时空特征, 紧接着的 LSTM 模块凭借其记忆功能将当前甚至更早时刻的信息传递给下一刻使用, 进一步提取时间特征, 同时采用时空注意力机制增强关键信息, 得到最终的 RGB 视频特征. ST-GCN 网络通过在骨骼序列上构建时空图并对其应用多层时空图卷积操作, 生成更高级的骨骼特征. 最后, 将提取到的两种模态特征融合, 采用标准的 Softmax 对融合特征向量做出动作分类预测. 下面详细介绍了算法中的每个模块.

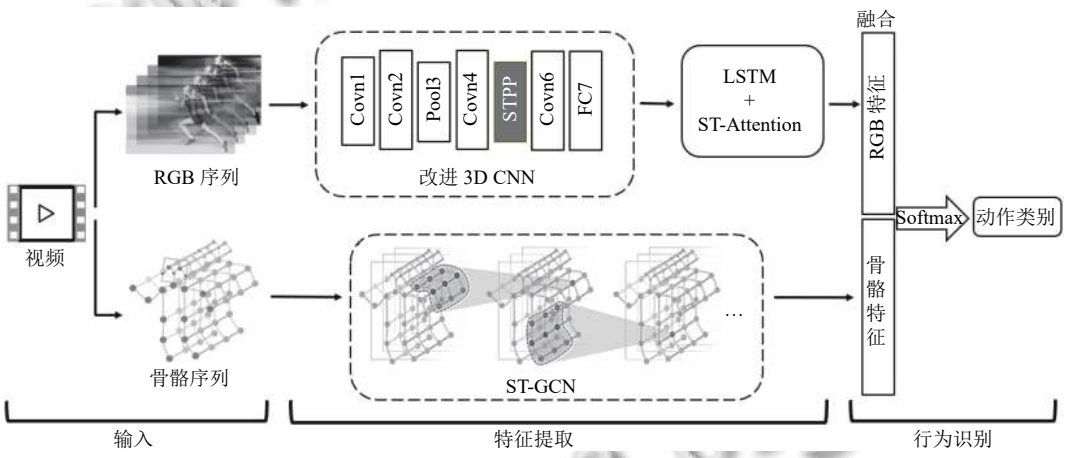

图 2 基于多模态特征学习的人体行为识别模型

2.1 改进 3D CNN 结构

3D 卷积网络与 2D 卷积网络相比, 可以同时捕捉目标的外观和运动信息, 具有更好的行为识别性能, 且其结构比目前许多行为识别模型都简单, 具有较快的运行速度. 该模型将多个连续帧堆叠形成立方体, 每帧中生成多个通道信息, 采用不同的核对连续帧的每一个通道做 3D 卷积, 这样得到的特征图连接到了多个邻近帧, 从而在提取空间信息的同时获得时间信息. 最终将所有通道上的信息结合计算得到多种类型的特征.

3D CNN 结构包括 1 个硬线层、3 个卷积层和 2 个下采样层, 网络以尺寸为 60×40 的连续 7 帧图像作为输入. 硬线层从每帧图像中提取灰度、横坐标梯度 x、纵坐标梯度 y、光流 x、光流 y 这 5 个通道信息, 生成 33 个特征图; C2 卷积层采用两种不同的 3D 核对上一层输出的 5 个通道信息分别进行卷积操作, C4 卷积层则采用 3 种不同的卷积核分别对特征图进行卷积操作, 从而得到更多的、兼具空间和时间两种维度的特征图; 降采样层 S3 和 S5 分别采用大小为 2×2 和 3×3 的滑动窗口对上一层得到的每个特征图进行下采样, 保持特征图数量不变的同时减少空间上的分辨率; 最后一个卷积层 C6 对每个特征图采用 7×4 的 2D 核进行卷积操作, 得到 128 个特征图, 即输入帧中动作信息的 128D 特征向量, 并送入全连接层做动作识别.

然而, 3D CNN 中全连接层的长度大小是事先定义好的, 这就要求网络的训练及测试都需要输入尺寸和比例固定的视频帧. 当输入任意大小的视频时, 3D

CNN 会对帧图像进行裁剪或缩放以产生固定大小的输入样本, 而这样操作很可能会导致重要信息丢失、扭曲, 从而影响特征的提取. 为了对任意尺寸的视频帧做更全面的处理, 本文用时空金字塔池化层替换掉 3D CNN 中最后一个池化层, 来接收大小不同的输入并将其转化为固定长度的特征向量, 同时提取更多不同时间角度的特征.

由于卷积层可以接收任意大小的输入, 并随之产生不同大小的输出. 给定一段任意尺寸的 RGB 视频序列作为 3D CNN 的输入, 经过前期的 3D 卷积和普通下采样后, 假设最后一个卷积层的特征映射尺寸为$T\times W\times H$, 其中 T 为池化立方体的时间, H 和W是帧的高度和宽度. 不同于 3D CNN 中使用的常规滑动窗口池化, STPP 在给定池化层产生的特征数量后, 会动态地调节滑动窗口的大小. 具体来说, 我们将$P(p_t,p_s)$表示为时空池化级, 其中p_t是时间池化级, p_s是空间池化级, 因此, 每个池化立方体的大小为$\lfloor T/p_t\rfloor\times\lfloor W/p_s\rfloor\times\lfloor H/p_s\rfloor$. 当$p_s=4,2,1$且$p_t=1$ 时, 大小不同的卷积输出就可转化为维度固定的特征向量, 输入全连接层. 其中, 每个时空池化立方体均对响应值采用最大池化. 这样, 配置了 STPP 的改进 3D CNN 就可以适应任意尺寸或比例的视频帧, 并支持对帧尺度的任意缩放.

2.2 基于时空注意力机制的 LSTM 模型

不同视频的长度不一定相同, 视频中每个动作的时间长度也是不一样的, 因为任何动作的发生都是一个动态的过程, 单纯的一帧视频图像或者连续几帧形成的片段常常不能在时间上表达出完整的动作. 然而, 3D CNN 只能接受长度固定 (7 帧) 的视频输入, 这导致任意长度视频的行为识别精度变低. 为了更充分地提取动作的连续特征, 本文在改进 3D CNN 后连接 LSTM 模型进一步识别人体行为.

LSTM 对输入或输出的长度没有固定限制, 这有利于捕捉任意长度数据的动作特征; 且作为循环神经网络的变体, 它不仅解决了 RNN 梯度爆炸的问题, 还对长期时间依赖关系具有很好的建模能力. LSTM 模块连接在改进 3D CNN 的全连接层后, 根据其特定的学习机制, 可以通过内部的门控单元对输入的数据选择性遗忘、记忆或更新, 获得可变长度的连续动作序列特征之间的关系. 此外, 由于时空注意力机制 (Spatial-Temporal Attention, ST-Att) 可以同时捕捉行为特征的空间相关性和动态时间相关性, 本文在 LSTM 模型中加入 ST-Att, 以筛选出权重较大的值, 增强关键特征, 获得更复杂的时空线索. 其单元结构如图 3 所示.

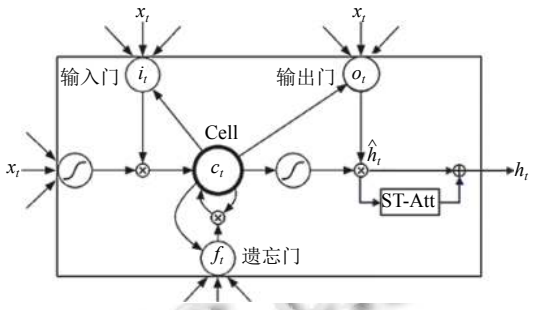

图 3 包含时空注意力机制的 LSTM 模型

在基本的 LSTM 单元中, i_t、f_t、o_t 分别代表 3 个门: 输入门, 遗忘门和输出门. i_t 根据传入信息选择性地更新细胞状态; f_t 负责对细胞状态中的信息选择性记忆或遗忘; o_t 控制的输出会对其他神经元产生一定的影响. c_t、$\hat{h}_t$、h_tc_t、$\hat{h}_t$、h_t则分别代表记忆细胞状态、LSTM 原始单元的输出和添加注意力后的输出. x_t代表行为视频经过改进 3D CNN 后得到的一系列特征, 具体计算公式如下:

$$\begin{cases} i_t=\sigma\left(W_{xi}x_t+W_{hi}\hat{h}_{t-1}+b_i\right)\\ f_t=\sigma\left(W_{xf}x_t+W_{hf}\hat{h}_{t-1}+b_f\right)\\ o_t=\sigma\left(W_{xo}x_t+W_{ho}\hat{h}_{t-1}+b_o\right)\\ g_t=\sigma\left(W_{xg}x_t+W_{hg}\hat{h}_{t-1}+b_g\right)\\ c_t=f_t\odot c_{t-1}+i_t\odot g_t\\ \hat{h}_t=o_t\odot\tanh(c_t)\\ h_t=f_{\text{att}}\left(\hat{h}_t\right)+\hat{h}_t \end{cases} \tag{1}$$

其中, $\sigma(\cdot)$表示取值范围为 (0, 1) 的 Sigmoid 非线性激活函数, $\tanh(\cdot)$ 表示取值范围为 (−1, 1) 的 tangent 非线性激活函, $\odot$表示哈达玛积, W_{ij}表示对应的权重矩阵, b_j代表偏差, $\hat{h}_{t-1}$表示隐藏状态, g_t 表示记忆调制状态, $f_{\text{att}}(\cdot)$表示能够自适应感知关键帧特征的注意力机制. 为了在加强关键帧信息的情况下不丢失非聚焦信息, 最终以$f_{\text{att}}\left(\hat{h}_t\right)$与$\hat{h}_t$的和作为输出, 保持时空特征的完整性.

2.3 时空图卷积网络

骨架序列能够有效地代表人体行为的动态, 目前, 我们已经可以通过 Kinect 和先进的人体姿态估计算法获得完整的 3D 骨架序列. 为了得到更加丰富的行为特征, 进一步提高动作识别精度, 本文采用 ST-GCN 作为基于骨架识别动作的通道模型. ST-GCN 是由图卷积网络扩展的时空图模型, 是用于行为识别的骨骼序列

通用表示, 它不仅拥有很强的表达能力和很高的性能, 而且易于在不同环境中推广.

首先, 对于具有 T 帧和 N 个节点的骨架序列, 构造表示该信息的时空图 $G=(V,E)$, 图中的节点集 $V=\{v_{ti}|t=1,\cdots,T;i=1,\cdots,N\}$ 表示骨骼序列中所有关节点, 每个节点都对应人体该处的关节, 其中第 t 帧的第 i 个节点的特征向量 $F(v_{ti})$ 由该节点的坐标向量和估计置信度组成. 这个图结构包括两种边: 一种是根据人体结构, 将每帧中的节点自然连接起来的空间边; 另一种是将连续两帧中相同关节点连接起来的时间边. 然后, 以构造的骨架图中关节点的坐标向量作为 ST-GCN 的输入, 对其应用多层时空图卷积操作, 图卷积后各关节的输出特征是由采样函数定义的邻域内关节特征的加权和, 最终得到人体行为视频的 3D 骨架特征图.

2.4 特征融合

对于行为识别, RGB 视频模态具有丰富的颜色和纹理信息, 3D 骨骼模态不容易受到光照、遮挡、衣着等不利因素的影响, 本文考虑到特征融合的互补性优势, 提出基于上述两种模态特征的人体行为识别方法. 根据融合发生的时间, 特征融合通常可分为: 早期融合、晚期融合和双向融合. 早期融合是指在进行识别之前将多种不同的特征融合, 其优势在于特征融合模块是独立于后期其他模型的. 因此, 本文采用早期融合策略, 将 RGB 视频和 3D 骨骼这两种类型的特征归一化后拼接起来, 生成新的混合特征向量, 并应用 Softmax 分类器对得到的融合特征进行动作分类. 融合后的特征可以使 RGB 视频与 3D 骨骼模态相辅相成, 优势互补, 从而传达重要的行为信息.

3 实验

3.1 数据集和评价标准

本文实验所用到的数据集为 UCF101[19] 和 NTU RGB+D[20]. UCF101 包含 13 320 个视频, 视频主要来源于 YouTube 等网站, 空间分辨率为 320×240. 该数据集共 101 个行为类别, 主要分为人和物体交互、只有肢体动作、人与人交互、玩音乐器材、各类运动五大类. 本文选取 9320 个视频用于训练, 剩下的 4000 个视频用于测试. NTU RGB+D 包含 56 880 个视频样本, 视频由 3 个 Microsoft Kinect v2 相机同时记录在不同水平视图下 40 个人的行为. 该数据集共有 60 个动作类别, 每个样本都包括 RGB 视频、深度图序列、3D 骨架数据和红外视频 4 种形式, RGB 视频的分辨率为 1920×1080, 深度图和红外视频均为 512×424, 3D 骨架数据包含每帧 25 个主要身体关节的三维位置. 本文选用 40 880 个视频作为训练集, 剩下 16 000 个视频作为测试集.

算法的评价标准为行为识别的准确率, 准确率取每个类别准确率的平均值.

3.2 训练细节

本文实验选择 Linux 操作系统和 PyTorch 深度学习框架. 首先, UCF101 数据集与 NTU RGB+D 数据集相比明显较小, 为了提高模型的泛化能力, 并且防止在 UCF101 上训练时出现过拟合现象, 本文对该数据集的视频做数据增广处理, 将样本扩充为原来的 5 倍. 其次, 为了减少视频长度对训练精度的影响, 统一将每个视频插值化处理为 32 帧. 由于本文在 3D CNN 中添加的 STPP 可以接受任意尺寸的输入, 因此不需要对两个数据集中视频的分辨率大小进行调整. 最后, 在特征融合阶段, 通过实验对比两种特征各占的权重, 选用 1:1.2 作为 RGB 特征和骨骼特征的权重.

训练时, 参考随机梯度下降算法中的参数, 将批处理大小设为 128, 动量设为 0.9. 将初始学习率设置为 0.001, 经过 15 000 次迭代后缩小 0.1, 最大迭代次数为 25 000 次.

3.3 实验结果与分析

本文的关键点主要在于: (1) 在 3D CNN 中添加 STPP; (2) 在提取 RGB 视频特征的通道加入包含时空注意力机制的 LSTM; (3) 将 RGB 特征与骨骼特征融合. 下面分别评估前两个模块对识别性能的影响, 并将最终识别模型与现有的流行方法做对比分析. 本文选用 UCF101 数据集, 添加各模块后的识别性能如表 1.

3.3.1 RGB 通道的模块分析

(1) 添加 STPP 的效果

带有 STPP 的改进 3D CNN 支持不同尺寸的视频输入而原始 3D CNN 不能, 因此, 本文在 UCF101 数据集上用多尺寸视频训练该模块, 用固定尺寸的视频训练 3D CNN. 由表 1 可知, 多尺寸训练的改进 3D CNN 比单尺寸训练的原始 3D CNN 效果要好, 识别精度提升了 2.4%, 这是因为多尺寸训练可以防止网络陷入过拟合.

(2) 添加基于时空注意力的 LSTM 的效果

由表 1 给出的在 UCF101 数据集上 LSTM 和时空注意力机制对视频行为的识别效果, 改进 3D CNN 连

接 LSTM 模型后的识别准确度有所提升; 进一步添加了时空注意力机制后, 性能优化更加明显, 准确度提高了 4.5%, 这是因为时空注意力机制可以有效地增强关键特征, 筛选出更复杂的时空信息, 从而提高模型的表达能力.

表 1 添加模块对识别性能的影响

算法	识别精度(%)
3D CNN	82.3
改进3D CNN	84.7
改进3D CNN+LSTM	85.4
改进3D CNN+LSTM+ST-Att	89.2

3.3.2 方法对比

将 C3DP-LA 和 ST-GCN 两个特征提取网络进行早期融合形成最终的识别模型, 为了评估模型性能, 本文将其与目前主流的深度学习算法进行比较.

(1) UCF101 数据集上的结果对比

表 2 给出了本文算法中 RGB 特征提取模型与双流卷积网络 (Two Stream), 3D 卷积网络 (3D CNN), 递归混合密度网络 (RMDN)[21], 时空注意力模型 (STA-CNN)[10] 的行为识别效果, 可以看出, 本文 RGB 通道模型的识别准确率优于其他算法, 表现出更好的性能.

表 2 不同行为识别算法在 UCF101 数据集上的准确率 (%)

算法名称	准确率
Two Stream[12]	88.0
3D CNN[8]	82.3
RMDN[21]	82.8
STA-CNN[10]	86.0
本文算法(仅RGB通道)	89.2

(2) NTU RGB+D 数据集上的结果对比

表 3 给出了本文最终识别模型与一些算法在交叉主体 (Cross-Subject, CS) 和交叉视图 (Cross-View, CV) 两个评估协议上的识别效果. 对比算法分为两类: 一类是基于单一模态 (如 RGB 或骨骼) 进行识别的模型, 包括姿态估计图的演化模型 (Pose Estimation Maps)[3], 关节轨迹图模型 (JTM)[7], 本文用到的时空图卷积网络 (ST-GCN) 和基于空间推理和时间堆栈学习的网络 (SR-TSL)[22]; 另一类是基于多种模态识别的模型, 包括手势识别网络 (STA-Hands)[23], 基于姿态的注意力模型 (Pose-based Attention)[24] 和深度聚合网络 (DAN)[25]. 由表 3 可以看出, 本文提出的方法在 NTU RGB+D 数据集上取得了 88.7% 和 92.8% 的识别准确率, 不仅优于单一模态的识别方法, 与其他多种模态融合的方法相比也表现出更好的性能, 证明了本文算法对人体行为识别的有效性.

表 3 不同行为识别算法在 NTU RGB+D 数据集上的准确率 (%)

算法名称	模态	CS	CV
Pose Estimation Maps[3]	RGB	78.8	84.2
JTM[7]	骨骼	73.4	75.2
ST-GCN[11]	骨骼	81.5	88.3
SR-TSL[22]	骨骼	84.8	92.4
STA-Hands[23]	RGB+骨骼	82.5	88.6
Pose-based Attention[24]	RGB+骨骼	84.8	90.6
DAN[25]	RGB+深度图	86.4	89.1
本文算法	RGB+骨骼	88.7	92.8

4 结论与展望

针对单一行为模态的特征难以充分表达复杂的人体动作, 导致行为识别准确度不高的问题, 本文提出基于多模态特征学习的行为识别算法, 分别学习视频的 RGB 特征和骨骼特征, 然后将两者融合, 利用融合特征的互补性优势, 达到提高行为识别率的目的. 通过在 UCF101 和 NTU RGB+D 两个公开的行为识别数据集上进行实验, 证明了本文方法与目前多种行为识别算法相比有着较高的识别准确率, 能够更有效地识别人体动作. 今后的研究将考虑到更多现实环境的因素, 提高算法实际应用时的在线识别精度和速度.

参考文献

1 Aggarwal JK, Ryoo MS. Human activity analysis: A review. ACM Computing Surveys, 2011, 43(3): 16. [doi: 10.1145/1922649.1922653]

2 Yeung S, Russakovsky O, Mori G, *et al*. End-to-end learning of action detection from frame glimpses in videos. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 2678–2687.

3 Liu MY, Yuan JS. Recognizing human actions as the evolution of pose estimation maps. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA. 2018. 1159–1168.

4 Weng JW, Weng CQ, Yuan JS, *et al*. Discriminative spatio-temporal pattern discovery for 3D action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(4): 1077–1089. [doi: 10.1109/TCSVT.2018.2818151]

5 Shotton J, Fitzgibbon A, Cook M, *et al*. Real-time human

pose recognition in parts from single depth images. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA. 2011. 1297–1304.
6 Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 588–595. [doi: 10.1109/CVPR.2014.82]
7 Wang PC, Li ZY, Hou YH, *et al*. Action recognition based on joint trajectory maps using convolutional neural networks. Proceedings of the 24th ACM International Conference on Multimedia Conference. Amsterdam, the Netherlands. 2016. 102–106.
8 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. 4489–4497.
9 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]
10 Meng LL, Zhao B, Chang B, *et al*. Interpretable spatio-temporal attention for video action recognition. arXiv: 1810.04511, 2018.
11 Yan SJ, Xiong YJ, Lin DH, *et al*. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, LA, USA. 2018. 7444–7452.
12 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 568–576.
13 Chen C, Liu K, Kehtarnavaz N. Real-time human action recognition based on depth motion maps. Journal of Real-time Image Processing, 2016, 12(1): 155–163. [doi: 10.1007/s11554-013-0370-1]
14 Lee I, Kim D, Kang S, *et al*. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. 2017. 1012–1020.
15 Chaaraoui AA, Padilla-López JR, Flórez-Revuelta F. Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. Proceedings of IEEE International Conference on Computer Vision Workshops. Sydney, NSW, Australia. 2013. 91–97. [doi: 10.1109/ICCVW.2013.19]
16 Sanchez-Riera J, Hua KL, Hsiao YS, *et al*. A comparative study of data fusion for RGB-D based visual recognition. Pattern Recognition Letters, 2016, 73: 1–6. [doi: 10.1016/j.patrec.2015.12.006]
17 Li M, Leung H, Shum HPH. Human action recognition via skeletal and depth based feature fusion. Proceedings of the 9th International Conference on Motion in Games. Burlingame, CA, USA. 2016. 123–132. [doi: 10.1145/2994258.2994268]
18 Chen C, Jafari R, Kehtarnavaz N. Improving human action recognition using fusion of depth camera and inertial sensors. IEEE Transactions on Human-Machine Systems, 2015, 45(1): 51–61. [doi: 10.1109/THMS.2014.2362520]
19 Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012.
20 Shahroudy A, Liu J, Ng TT, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. 2016. 1010–1019.
21 Bazzani L, Larochelle H, Torresani L. Recurrent mixture density network for spatiotemporal visual attention. arXiv: 1603.08199, 2016.
22 Si CY, Jing Y, Wang W, *et al*. Skeleton-based action recognition with spatial reasoning and temporal stack learning. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2019. 106–121.
23 Baradel F, Wolf C, Mille J. Human action recognition: Pose-based attention draws focus to hands. Proceedings of 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy. 2017. 604–613. [doi: 10.1109/ICCVW.2017.77]
24 Baradel F, Wolf C, Mille J. Human activity recognition with pose-driven attention to RGB. Proceedings of the 29th British Machine Vision Conference (BMVC). Newcastle, UK. 2018. 1–14.
25 Wang PC, Li WQ, Wan J, *et al*. Cooperative training of deep aggregation networks for RGB-D action recognition. Proceedings of 32nd AAAI Conference on Artificial Intelligence(AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, LA, USA. 2018. 7404–7411.