计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):41−49 [doi: 10.15888/j.cnki.csa.008805]

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

基于深度学习的行为识别多模态融合方法综述①

詹健浩[1], 吴鸿伟[2], 周成祖[2], 陈晓筹[3], 李晓潮[1]

[1](厦门大学 电子科学与技术学院, 厦门 361005)
[2](厦门市美亚柏科信息股份有限公司, 厦门 361016)
[3](厦门大学 信息与网络中心, 厦门 361005)
通信作者: 李晓潮, E-mail: leexcjeffrey@xmu.edu.cn

摘　要: 行为识别是通过对视频数据进行处理分析从而让计算机理解人的动作和行为. 不同模态数据在外观、姿态、几何、光照和视角等主要特征上各有优势, 通过多模态融合将这些特征进行融合可以获得比单一模态数据更好的识别效果. 本文对现有行为识别多模态融合方法进行介绍, 对比了它们之间的特点以及获得的性能提升, 包括预测分数融合、注意力机制、知识蒸馏等晚期融合方法, 以及特征图融合、卷积、融合结构搜索、注意力机制等早期融合方法. 通过这些分析和比较归纳出未来多模态融合的研究方向.

关键词: 行为识别; 深度学习; 多模态融合; 晚期融合; 早期融合

引用格式: 詹健浩,吴鸿伟,周成祖,陈晓筹,李晓潮.基于深度学习的行为识别多模态融合方法综述.计算机系统应用,2023,32(1):41–49. http://www.c-s-a.org.cn/1003-3254/8805.html

Survey on Multi-modality Fusion Methods for Action Recognition Based on Deep Learning

ZHAN Jian-Hao[1], WU Hong-Wei[2], ZHOU Cheng-Zu[2], CHEN Xiao-Chou[3], LI Xiao-Chao[1]

[1](Department of Microelectronics and Integrated Circuit, Xiamen University, Xiamen 361005, China)
[2](Xiamen Meiya Pico Information Co. Ltd., Xiamen 361016, China)
[3](Information and Network Center, Xiamen University, Xiamen 361005, China)

Abstract: Action recognition aims to make computers understand human actions by the processing and analysis of video data. As different modality data have different strengths in the main features such as appearance, gesture, geometric shapes, illumination, and viewpoints, action recognition based on the multi-modality fusion of these features can achieve better performance than the recognition based on single modality data. In this study, a comprehensive survey of multi-modality fusion methods for action recognition is given, and their characteristics and performance improvements are compared. These methods are divided into the late fusion methods and the early fusion methods, where the former includes prediction score fusion, attention mechanisms, and knowledge distillation, and the latter includes feature map fusion, convolution, fusion architecture search, and attention mechanisms. Upon the above analysis and comparison, the future research directions are discussed.

Key words: action recognition; deep learning; multi-modality fusion; late fusion; early fusion

基于深度学习的行为识别是通过对视频数据进行处理分析从而让计算机理解人的动作和行为, 是计算机视觉领域中一个重要的研究领域, 在视频监控[1]、自主导航[2]、视频检索[3]等方面有着广泛的应用. 行为识别领域论文近期在各顶会接收数目如表 1 所示, 从中可以看出近年来行为识别领域接收的论文占论文总接

① 基金项目: 福建省高校产学研联合创新项目 (2022H6004); 集成电路设计与测试分析福建省高校重点实验室基金; 厦门大学马来西亚研究基金 (XMU-MRF/2019-C4/IECE/0008)
收稿时间: 2022-03-08; 修改时间: 2022-04-12; 采用时间: 2022-04-20; csa 在线出版时间: 2022-07-14
CNKI 网络首发时间: 2022-11-15

收数的比例约在 3%–4% 左右, 在 AAAI 会议中还有逐年增长的趋势, 说明了行为识别领域有着深远的研究意义和研究价值.

早期行为识别的研究主要集中单一模态数据的分析上, 由于不同的单一模态数据有自己的优势与不足, 如: RGB 模态含丰富的外观信息但对遮挡、环境变化或阴影的干扰敏感[4] 等. 多模态数据包含场景中不同的特征信息, 不同模态数据在外观、姿态、几何、光照和视角等主要特征上各有优势. 因此, 如何结合多个模态数据各自的优势从而获得更好的识别效果得到了越来越多的关注, 行为识别任务也正在从原先的单一模态行为识别向多模态行为识别发展. 多模态行为识别是从多模态数据中提取和组合相关信息, 获得比单一模态数据更加丰富、互补的信息, 从而得到比只使用单一模态更好的识别性能[5,6]. 多模态数据是从不同来源或传感器收集的异构数据, 不同模态各有其优缺点并可以互补[7], 如: RGB 模态外观信息丰富但受光照、背景变化影响大, 而骨骼模态对背景变化鲁棒且关注人体骨骼运动但缺乏人与物体的外观信息[8]. 因此, 由于多模态行为识别所具有的独特优势使得多模态行为识别方法受到了广泛的关注和研究[4–6,8].

表 1 行为识别论文在各顶会的接收数 (接收数/总接收数)

年份	CVPR	ICCV	ECCV	AAAI
2019	42/1 300	33/1 077	—	24/1 150
2020	50/1 470	—	44/1 361	27/1 591
2021	43/1 663	—	—	43/1 692

注: ICCV和ECCV每两年召开一次.

多模态行为识别的关键和难点在于如何从不同的模态数据中提取和组合相关互补信息, 即多模态融合[9–24]. 多模态数据模态分为视觉模态和非视觉模态[7], 视觉模态主要有 RGB、光流、骨骼、红外、深度、点云等模态, 而非视觉模态主要有音频、雷达、WiFi 模态等, 本文主要研究视觉模态的多模态融合. 多模态融合的目的在于融合多种不同模态之间提取的互补语义信息来学习有效的特征从而提高性能[24], 主要分为早期融合 (early fusion) 和晚期融合 (late fusion)[25]. 早期融合是将不同模态提取的特征进行融合, 即特征融合 (feature fusion)[24]; 而晚期融合是将单一模态的预测分数进行融合, 即分数融合 (score/decision fusion)[24]. 早期融合常见方法有: 特征图相加[9]、特征图堆叠[9,10]、卷积[11,23]、多模态融合结构搜索 (multimodal fusion architecture search)[13]、注意力机制 (attention mechanism)[14,15] 等. 晚期融合通常采用预测分数取平均值[10,16–18,26,27]、最大值[16,18]、相乘[16,18]、相加[16]、注意力机制[16]、知识蒸馏[20,26] 等方式. 目前的研究中就性能而言, 行为识别的多模态晚期融合相对于单一模态性能往往能获得一定的提升[15,18,23,26,27], 但晚期融合忽略了多模态特征之间的相关性[28], 获得的提升有限.

由于早期融合能够融合不同层次的语义信息, 在最新的研究中早期融合比晚期融合获得了更好的效果[13–15,23]. 尽管晚期融合方法简单, 但晚期融合是将每个单一模态网络最后一层得到的预测分数进行融合, 而预测分数仅表示网络抽象的语义预测, 缺少不同模态间中、低层次特征的融合, 而这方面正是早期融合的重点. 网络不同层提取的特征代表着不同层次的语义信息, 因此如何将不同模态提取的不同层次的语义信息进行融合是早期融合的技术关键. 在早期研究中, 采用特征图相加、特征图堆叠[9] 等简单方法进行融合由于需要额外的卷积层进行建模, 加上数据集规模偏小无法充分地对多模态信息进行建模和融合[9], 使得早期融合取得的效果并不如晚期融合. 为了解决这个问题, 将有着有效信息建模能力的注意力机制用于多模态信息提取[15]、将同时提取及融合外观、运动信息的 SlowFast 网络结构用于多模态早期融合[23] 等方法的提出, 使早期融合的性能提升超过了晚期融合, 并获得了越来越多的关注.

如何有效提取不同模态独有的信息和如何选择不同层次的语义信息进行融合[13,22] 是多模态早期融合中的技术关键. 一方面, 不同模态含有各自独有的对行为识别分类有帮助的信息[15,23], 如: 骨骼模态对视角、光照变化不敏感且提供人体动作的关键几何信息但缺少外观特征, 而 RGB 模态有着丰富的外观特征[14], 因此如何更好地在模型不同卷积层中提取这些信息用于多模态早期融合十分重要. 另一方面, 由于卷积神经网络随着卷积层的加深, 提取的语义信息越来越抽象, 如低层的卷积层往往作为边缘检测器提取低层次的语义信息, 而高层的卷积层提取的是更抽象复杂的语义信息, 因此如何选择不同层次的语义信息进行融合是十分困难的[13], 而在多模态情况下这个问题尤为突出: 首先, 以图的形式对 3D 骨骼模态进行建模的图卷积神经网络难以与传统的卷积神经网络进行融合[23]; 其次, 由于不同模态含有的信息丰富程度不同, 如 RGB 模态较骨

骼模态含更丰富的信息, 因此骨骼模态不需要建模 RGB 模态那么复杂的网络进而避免过拟合[10], 网络层数的差异导致不同模态提取出的语义信息也不处于同一层次, 进而使得特征融合更加艰难.

综上所述, 多模态晚期融合是对训练完成的多个模态网络预测分数的融合, 如表 2 所示, 其中 CS 指标按照人物来划分训练集和测试集, CV 指标按相机来划分训练集和测试集, 可以看出近期 RGB+光流的双流晚期融合方法在常用的行为识别数据集 UCF101 和 HMDB51 上分别获得了 1%–2% 和 1%–4% 的提升, 对 RGB、2D/3D 骨骼及深度等模态进行晚期融合在常用多模态行为识别数据集 NTU RGB+D 60 上普遍可以获得 1%–3% 的提升, 但由于晚期融合忽略了不同模态特征之间的相关性[28] 阻碍了性能的进一步提升.

表 2 近期研究中多模态晚期融合及其性能提升效果对比

时间	方法	模态	数据集	准确率(性能提升) (%)
2019	MARS[26]	RGB+光流	UCF101 HMDB51	97.1 (1.9) 80.1 (4.2)
2019	LGD-3D[27]	RGB+光流	UCF101 HMDB51	98.2 (1.2) 80.5 (1.6)
2020	Inflated ResNet50[29]	RGB+2D骨骼	NTU RGB+D 60	CS: 95.66 (0.41), CV: 98.79 (0.45)
2020	MMTM[15]	RGB+2D骨骼	NTU RGB+D 60	CS: 91.56 (2.31)
2021	BPAN[16]	RGB+3D骨骼	NTU RGB+D 60	CS: 93.07 (1.5)
2021	Res3D-101[18]	深度+2D骨骼	NTU RGB+D 60	CS: 91.13 (2.04), CV: 94.31 (2.53)
2021	RGBPose-SlowFast[23]	RGB+2D骨骼	NTU RGB+D 60	CS: 95.5 (1.4)

多模态早期融合是对不同模态特征的融合, 但由于有效提取不同模态独有的信息和选择不同层次的语义信息进行融合[13,22] 较为困难影响性能的提升, 但相关研究仍有广阔的前景, 如表 3 所示, 其中, NTU RGB+D 系列数据集中的 C-Sub 指标按照人物来划分训练集和测试集, C-Set 指标按相机的设置号来划分训练集和测试集, 最近研究较多的对 RGB 和 2D/3D 骨骼模态在 NTU RGB+D 60 数据集上进行早期融合可以获得 2%–5% 的性能提升, 获得的提升已经超过晚期融合方法, 在 NTU RGB+D 120 数据集上可以获得 7%–10% 的性能提升, 由此可见随着对早期融合方法的研究深入, 其对性能的提升越来越明显, 早期融合或将成为多模态融合未来主要的研究方向.

本文的主要贡献如下所示:

(1) 本文详细介绍了多模态融合中早期融合和晚期融合的各种方法的原理及其获得的性能提升.

(2) 本文对近期多模态晚期融合和早期融合的研究做了综合地对比, 分析了它们的识别效果和性能提升. 通过这些分析和比较归纳出早期融合或将成为多模态融合未来主要的研究方向.

表 3 近期研究中多模态早期融合及其性能提升效果对比

时间	方法	模态	数据集	准确率 (性能提升) (%)
2019	MFAS[13]	RGB+3D骨骼	NTU RGB+D 60	CS: 90.04 (4.80)
2020	VPN[14]	RGB+3D骨骼	NTU RGB+D 60 NTU RGB+D 120	CS: 95.5 (2.5), CV: 98.0 (2.6) C-Sub: 86.3 (9.3), C-Set: 87.8 (7.7)
2020	MMTM[15]	RGB+2D骨骼	NTU RGB+D 60	CS: 91.99 (2.74)
2021	RGBPose-SlowFast[23]	RGB+2D骨骼	NTU RGB+D 60	CS: 96.2 (2.1)

本文第 1 节介绍了多模态行为识别常用的多模态数据, 第 2 节介绍了多模态晚期融合和早期融合的过程, 第 3 节详细介绍了最近多模态融合研究中关于晚期融合的各种方法, 第 4 节详细介绍了最近多模态融合研究中关于早期融合的各种方法, 第 5 节对晚期融合和早期融合进行了对比研究并讨论了两种融合的效果和趋势, 第 6 节对多模态融合方法做了总结和展望.

1 多模态行为识别常用的多模态数据

1.1 多模态数据及其特点

行为识别领域主要使用的多模态数据有: RGB、光流、骨骼 (3D/2D)、深度、红外等. 其中, 每一种模态数据都有自己的特点[7], 如表 4 所示, 其中各模态在外观信息、获取难度、运动信息、几何信息、光照、视角和背景变化等方面各有优势和不足, 如其中由于

2D/3D 骨骼模态信息对视角、背景变化不敏感, 可以弥补 RGB 模态对视角背景敏感的缺点, 使得表 2 和表 3 中的 RGB 和 2D/3D 骨骼模态融合广泛应用于 NTU RGB+D 系列多模态数据集中; 由于光流模态含有丰富的运动信息, 而 RGB 模态含有丰富的外观信息可以弥补光流模态外观信息不足的缺点, 使得 RGB 和光流的双流晚期融合广泛用于 UCF101 和 HMDB51 数据集中. 由此可见, 不同模态数据可以提取出互补的特征信息, 通过多模态融合将这些互补的信息进行融合可以获得比单一模态数据更好的识别效果.

表 4 多模态数据特点表

模态	优点	缺点
RGB	外观信息丰富	对光照、视角、遮挡、背景变化敏感
光流	运动信息丰富	需要额外的计算无法实时获取、缺乏外观信息
2D/3D 骨骼	对视角、背景等变化不敏感	缺乏外观信息、对遮挡敏感
深度	人体几何信息和3D结构信息丰富, 对视角、色彩变化不敏感	缺少色彩和背景信息、可工作距离受限
红外	不受光照影响	缺乏外观和背景信息、易受温度影响

1.2 多模态数据的输入形式

目前多模态行为识别任务常用的多模态数据主要有两种输入形式, 一种是使用提供多种模态数据的多模态数据集[5,14–16], 而另一种是经转换的数据模态, 如: 采用算法从 RGB 模态提取骨骼模态等[8,10,18,20,23,30]. 其中骨骼模态分为 3D 骨骼模态和 2D 骨骼模态 (也称为姿态模态), 3D 骨骼模态主要以关节点坐标的形式保存, 2D 骨骼模态主要以图像形式保存.

1.2.1 多模态数据集

目前最常用的行为识别多模态数据集是 NTU RGB+D 系列数据集 (包含: NTU RGB+D 60[31] 和 NTU RGB+D 120[32]). 其中, NTU RGB+D 系列数据集由 3 台 Kinect 相机采集, 其中 NTU RGB+D 60 含 60 类动作包含 56 880 个样本、NTU RGB+D 120 包含 120 类动作和 114 480 个样本, 并都含有 40 个受测者的 RGB 模态、深度模态、3D 骨骼模态和红外模态信息, NTU RGB+D 60 有两个性能评价标准: CS (cross-subject) 和 CV (cross-view), CS 是按照人物来划分训练集和测试集, 将人物 ID 特定的 20 人作为训练集, 剩余的作为测试集, 而 CV 是按相机来划分训练集和测试集, 相机 2 和 3 采集的样本作为训练集, 相机 1 采集的样本作为测试集; NTU RGB+D 120 有两个性能评价标准: C-Sub (cross-subject) 和 C-Set (cross-setup), C-Sub 同样按照人物 ID 来划分训练集和测试集, C-Set 按照相机的设置号划分训练集和测试集, 将偶数设置号的受测者的 63 026 个样本视频作为训练集, 奇数设置的 50 919 个样本作测试集. 另外, 常用的仅含 RGB模态的行为识别数据集有 UCF101[33] 和 HMDB51 数据集[34], UCF101 共计 13 320 个视频总计 101 个动作类别, 而 HMDB51 的视频大多数来源于互联网和电影, 受光照和视角变化、背景遮挡等因素影响较大, 共计 6 849 个视频数据, 总计 51 个动作类别.

1.2.2 经转换的数据模态

使用经转换的数据模态作为数据输入的方法, 主要有: 文献 [26,27,35] 等将 UCF101 和 HMDB51 数据集使用稠密光流 TV-L1 算法获取 RGB 模态时间维度上相邻帧之间的变化来计算运动信息, 从而将 RGB 模态提取为光流模态; Bian 等人[30] 讨论了采用 Openpose[36] 和 PifPaf[37] 将 RGB 模态转化为 2D 骨骼模态的效果; Yan 等人[10] 和 Wu 等人[18] 同样采用了 Openpose[36] 将 RGB 模态转化为 3D 骨骼模态; Li 等人[20] 采用了 PoseNet[38] 将 RGB 模态转化为 2D 骨骼模态; Duan 等人[23] 采用 HRNet[39] 将 RGB 模态转化为 2D 骨骼模态; Das 等人[40] 通过使用 LCR-Net++[41] 将 RGB 模态转化为 2D 骨骼模态.

2 多模态晚期融合和早期融合的过程

多模态融合方法包括多模态晚期融合和早期融合, 在多模态融合过程中, 多模态数据输入为: $D=\left\{\left(x_i^p,y_i\right)\mid 1\leqslant p\leqslant P,1\leqslant i\leqslant n\right\}$[28], D是包含P个模态共n个样本的数据输入, x_i^p表示第p个模态的第i个数据样本对应的特征向量, y_i表示第i个样本的标签. 多模态晚期融合的过程可以表示为[28]:

$$\mathcal{F}\left(\left\{\min\frac{1}{n}\sum_{i=1}^{n}\ell\left(h_p\left(x_i^p\right),y_i\right)\right\}_{p=1}^{P}\right) \tag{1}$$

其中, x_i^p表示样本i的第p ($1\leqslant p\leqslant P$)个模态对应的特征向量, h^p表示样本输入i模态p对应神经网络所输出的预测结果, y_i表示样本i对应的标签, ℓ表示求交叉熵损失函数, $\mathcal{F}$多模态融合操作. 从式 (1) 可以看出, 多模态晚期融合是先训练多个模态对应的网络, 然后将训练好的各模态网络的预测分数进行融合, 属于分数融合.

多模态早期融合的过程可以由式 (2) 表示[28]:

$$\min \frac{1}{n}\sum_{i=1}^{n}\ell\left(h\left(\mathcal{F}\left(x_i^1,x_i^2,\cdots,x_i^P\right)\right),y_i\right) \tag{2}$$

其中, x_i^p表示样本i的第p ($1\leqslant p\leqslant P$)个模态对应的特征向量, $\mathcal{F}$表示多模态融合操作, h表示融合后的特征向量输入神经网络后的预测分数, y_i表示输入样本i对应的标签, ℓ表示对预测分数和标签求交叉熵损失函数. 从式 (2) 可以看出, 模态早期融合是先融合多个模态的特征, 然后将融合后的特征输入网络并对网络进行训练, 属于特征融合.

3 晚期融合

行为识别多模态的晚期融合是指将每个模态数据分别用网络对其建模, 每个模态对应网络最后一层输出会得到该模态的预测分数, 将得到的预测分数采用某种方式进行融合即多模态晚期融合, 目的是融合基于不同模态分别做出的决策, 通常采用预测分数取平均值[10,16–18,26,27]、最大值[16,18]、相乘[16,18]、相加[16]、注意力机制[16]、知识蒸馏[20,26] 等方式, 表 5 描述了多种多模态晚期融合方法及其效果, 可以看出采用预测分数取平均值方法进行晚期融合在多模态行为识别数据集 NTU RGB+D 60 中融合 RGB、光流、深度和 2D/3D 骨骼模态时均有 1%–3% 的提升, 在 UCF101 和 HMDB51 上同样也有 1%–4% 的提升; 采用预测分数取最大值和预测分数相乘的方法进行晚期融合的效果不如同等情况下采用预测分数取平均值的效果; 采用预测分数相加的方法与采用预测分数取平均值的方法获得的性能提升几乎相等; 采用注意力机制的方法进行效果目前最好在 NTU RGB+D 60 数据集上达到了 3.85% 的提升; 采用知识蒸馏的方法进行晚期融合在 UCF101 和 HMDB51 上获得了 1%–4% 左右的提升.

表 5 多模态晚期融合方法及其性能提升效果对比

方法	文献	时间	模态	数据集	准确率 (性能提升) (%)
预测分数取平均值	[16]	2021	RGB+3D骨骼	NTU RGB+D 60	CS: 93.07 (2.07)
	[18]	2021	深度+2D骨骼	NTU RGB+D 60	CS: 91.13 (2.04), CV: 94.31 (2.53)
	[23]	2021	RGB+2D骨骼	NTU RGB+D 60	CS: 95.5 (1.4)
	[26]	2019	RGB+光流	UCF101 HMDB51	97.5 (1.7) 79.8 (3.9)
	[27]	2019	RGB+光流	UCF101 HMDB51	98.2 (1.2) 80.5 (1.6)
预测分数取最大值	[16]	2021	RGB+3D骨骼	NTU RGB+D 60	CS: 92.35 (1.35)
	[18]	2021	深度+2D骨骼	NTU RGB+D 60	CS: 89.06 (0.03), CV: 91.97 (0.19)
预测分数相乘	[16]	2021	RGB+3D骨骼	NTU RGB+D 60	CS: 92.20 (1.20)
	[18]	2021	深度+2D骨骼	NTU RGB+D 60	CS: 90.77 (1.68), CV: 94.30 (2.52)
预测分数相加	[16]	2021	RGB+3D骨骼	NTU RGB+D 60	CS: 93.09 (2.09)
注意力机制	[16]	2021	RGB+3D骨骼	NTU RGB+D 60	CS: 94.85 (3.85)
知识蒸馏	[26]	2019	RGB+光流	UCF101 HMDB51	97.1 (1.3) 80.1 (4.2)
	[42]	2020	RGB+光流	UCF101 HMDB51	97.6 (0.8) 80.5 (3.2)

预测分数取平均值的方法是多模态晚期融合中最常见的融合方法, 其中行为识别中传统的 RGB+光流的双流结构[17,26,27,35] 对 RGB 模态和光流模态对应模型单独训练, 测试使对两个网络的预测分数取平均值来实现晚期融合. Wu 等人[18] 研究深度模态和 2D 骨骼模态, 其中对深度模态计算每帧间的变化信息来表示运动信息并采样输入 3D 卷积神经网络, 对 RGB 模态输入提取对应的 2D 骨骼模态并通过采样输入 3D 卷积神经网络, 最后通过对预测分数取平均值得到了晚期融合中最好的效果. Xu 等人[16] 和 Wu 等人[18] 同时还探索了采用取不同模态预测分数的最大值、相乘以及相加的方法进行多模态晚期融合, 表 5 中的实验效果表明这些方法的性能提升不如预测分数取平均值的方法.

注意力机制融合预测分数的方法在最近的研究中得到很好的效果, Xu 等人[16] 设计了 BPAN 注意力模块将网络提取的 RGB 和 2D 骨骼模态的预测分数求矩阵外积后输入注意力模块得到注意力权值, 将注意力权值与原 RGB 和 2D 骨骼模态的预测分数相乘并对其

分别求损失函数进行训练.

知识蒸馏的方法旨在将学生网络向教师网络做回归[28], 将知识从教师网络转移到学生网络[30]. Crasto 等人[26] 提出了使用知识蒸馏的方法, 将训练好的光流模态网络作为教师网络将 RGB 模态网络作为学生网络进行训练, 获得了比任一单模态网络更高的精度. Li 等人[20] 提出了一个多教师知识蒸馏的方法, 先训练光流和 2D 骨骼模态作为教师网络, 而后训练 RGB 模态学生网络时将两个教师网络的预测分数与学生网络预测分数做均方误差损失函数, 即 RGB 模态的学生网络在预测分数层面对教师网络进行回归从而实现晚期融合. 与其他晚期融合方法不同的是, 知识蒸馏方法虽然训练时也需要多模态数据, 但测试仅需学生网络以及相应模态数据进行测试即可.

4 早期融合

早期融合是指将不同模态卷积神经网络提取的特征图进行融合, 常见的方法有: 特征图相加[9]、特征图堆叠[9,10]、卷积[11,23]、多模态融合结构搜索[13]、注意力机制[14,15,43] 等, 表 6 描述多模态早期融合方法及其效果, 采用简单的特征图相加和特征图堆叠方法进行早期融合提升很少甚至有性能下降的现象; 采用卷积的方法在 NTU RGB+D 60 和 UCF101 数据集上融合 RGB、光流、骨骼等模态可以获得约 2% 的提升; 采用多模态融合结构搜索的方法进行早期融合可以获得约 5% 的提升; 采用注意力机制的方法进行早期融合在 NTU RGB+D 60 数据集上可以获得 1%–3% 的提升, 在 NTU RGB+D 120 数据集上可以获得 7%–10% 的提升. 对比表 5 和表 6 在最新研究[23]中相同情况下早期融合和晚期融合的效果, 可以看出随着越来越有效的早期融合方法的提出, 早期融合可以更有效地利用网络提取的不同模态的不同层次的语义信息, 早期融合带来的性能提升已经超过了晚期融合, 早期融合或将成为多模态融合未来主要的研究方向.

表 6 多模态早期融合方法及其性能提升效果对比

方法	文献	时间	模态	数据集	准确率 (性能提升) (%)
特征图相加	[9]	2020	RGB+2D骨骼	Volleyball	91.2 (–0.3)
特征图堆叠	[9]	2020	RGB+2D骨骼	Volleyball	91.8 (0.3)
	[10]	2019	RGB+2D骨骼	JHMDB	58.7 (–1.4)
卷积	[11]	2019	RGB+光流	JHMDB	60.99 (3.0)
	[23]	2021	RGB+2D骨骼	NTU RGB+D 60	CS: 96.2 (2.1)
多模态融合结构搜索	[13]	2019	RGB+2D骨骼	NTU RGB+D 60	CS: 90.04 (4.8)
注意力机制	[14]	2020	RGB+3D骨骼	NTU RGB+D 120	C-Sub: 86.3 (9.3), C-Set: 87.8 (7.7)
	[15]	2020	RGB+2D骨骼	NTU RGB+D 60	CS: 91.99 (2.74)
	[43]	2021	RGB+2D骨骼	NTU RGB+D 60	CS: 91.7 (1.3)

特征图相加和特征图堆叠的方法是早期融合中较为简单常见的融合方法[9,10]. Gavrilyuk 等人[9] 主要研究群体行为, 将多模态信息分为包含人物边框信息的 RGB 模态图像序列、光流模态图像序列以及单张 2D 骨骼模态图像, 然后分别将 RGB 模态和光流模态的图像序列输入到 3D 卷积神经网络、将单张 2D 骨骼模态图像输入到 2D 网络, 融合各自得到的特征图并输入 Transformer 和剩下的网络, 得到最终的预测结果. 文献 [9] 中采用的早期融合讨论了特征图相加、特征图堆叠, 特征图相加并未得到性能上的提升, 而特征图堆叠获得了一定的提升. 文献 [10] 中的输入含 RGB 和 2D 骨骼模态, 通过姿态网络将 RGB 模态转化为 2D 骨骼热图, 热图包含 3 部分: 关节、部位和特征热图, 将这 3 部分热图进行特征堆叠完成早期融合并输入 3D 卷积神经网络得到骨骼模态的预测结果, 最后将两个模态预测结果融合得到最终的预测结果.

卷积的方法是早期融合中提取不同模态信息的有效方法[11,23]. Zhao 等人[11] 将光流模态输入网络得到特征图后, 将光流特征图与 RGB 模态网络中不同层的特征图进行多次融合, 融合方式为: 先将光流特征通过第一个卷积并与 RGB 特征图按位乘得到第 1 步融合后的特征图, 再将光流特征通过第 2 个卷积与第 1 步融合后的特征图按位加来实现早期融合. Duan 等人[23] 使用了 SlowFast 网络结构, RGB 模态使用模型复杂度较高的 Slow 网络来建模, 而 2D 骨骼模态是先通过对 RGB 模态进行人物检测并裁剪后使用姿态估计网络提取对应的 2D 骨骼热图, 并输入到模型复杂度较低的 Fast 网络来建模, 其中 2D 骨骼模态的 Fast 网络的特征

通过卷积后与 Slow 网络特征图按位加、特征图堆叠的方式与 RGB 模态进行融合.

多模态融合结构搜索[13] 的方法是为了解决如何选择不同层次的语义信息进行多模态融合的问题. Pérez-Rúa 等人[13] 提出了一个基于神经结构搜索[39] 的多模态融合结构搜索算法的方法, 通过网络训练来解决两个模态间如何选取两个网络不同层特征图进行融合以及如何选用更合适的非线性函数的问题.

注意力机制[14,15,43] 进行早期融合的方法随着注意力机制的发展得到了越来越多的关注. Das 等人[14] 对 RGB 模态采用 3D 卷积神经网络提取特征, 对 3D 骨骼模态采用图卷积神经网络提取骨骼信息并使用注意力机制对其进行信息提取得到注意力权值, 将权值与 RGB 模态特征矩阵乘并与原 RGB 模态特征矩阵加得到融合后的特征, 同时对 RGB 模态特征与骨骼模态求损失函数来保证两个模态时空信息的耦合, 保证多模态融合效果. Joze 等人[15] 参照 SENet[38] 扩展了用于 3DCNN 的 MMTM 注意力模块并将 RGB 模态和 2D 骨骼模态的两层特征图作为注意力模块的输入得到注意力权值并与各模态输入特征图矩阵乘进行早期融合, 在两个模态网络不同层进行重复多次该早期融合操作, 最后在预测分数进行矩阵按位加得到融合后预测分数进行预测, 获得了很好的效果. Moon 等人[43] 使用 2D 卷积神经网络和时间卷积对 RGB 和 2D 骨骼模态分别进行建模, 对 2D 骨骼模态使用注意力机制产生门矩阵, 然后将门矩阵与 RGB 模态特征矩阵乘、将单位矩阵与门矩阵差与 2D 骨骼模态特征进行矩阵乘操作调整, 最后将两个调整后的特征按位加实现融合.

5 晚期融合和早期融合的对比研究

本节主要将近期晚期融合和早期融合在最常用的多模态数据集 NTU RGB+D 60 上的识别准确率和性能提升进行一个横向的对比和讨论, 如表 7 所示, 并对比了晚期融合和早期融合的性能提升, 接此寻找晚期融合和早期融合的一些发展规律.

总的来说, 现阶段多模态早期融合与晚期融合相比可以达到更好的效果. 表 7 晚期融合中的文献 [16] 和文献 [23] 的 CS 指标的识别准确率分别达到了 94.85% 和 95.5%, 在目前晚期融合研究中处于较高水平; 而早期融合中的文献 [14,23,29] CS 指标的识别准确率分别达到了 95.5%、95.45% 和 96.2%, 在早期融合研究中处于较高水平. 其中, 文献 [16] 采用晚期融合方法融合 RGB 和 3D 骨骼模态获得 94.85% 的准确率, 而文献 [14] 采用早期融合方法同样融合 RGB 和 3D 骨骼模态获得了 95.5% 的准确率. 文献 [23] 同时做了晚期融合和早期融合的实验, 采用早期融合方法融合 RGB 和骨骼模态并取得 96.2% 的准确率, 比相同情况下采用晚期融合的 95.5% 进一步提升了 0.7% 的识别准确率, 进一步证明了早期融合的有效性. 表 7 中采用早期融合的文献 [15,43,44,45] 等识别准确率在 91%–92% 左右, 而采用晚期融合的文献 [5,18] 等识别准确率在 90%–91% 左右.

表 7 多模态晚期融合和早期融合在 NTU RGB+D 60 数据集上的性能对比

方法	文献	时间	模态	准确率 (%)
晚期融合	[16]	2021	RGB+3D骨骼	CS: 94.85
	[5]	2021	RGB+深度+光流	CS: 89.70, CV: 92.97
	[18]	2021	深度+2D骨骼	CS: 91.13, CV: 94.31
	[23]	2022	RGB+2D骨骼	CS: 95.5
早期融合	[44]	2018	2D骨骼+3D骨骼	CS: 91.71, CV: 95.26
	[13]	2019	RGB+2D骨骼	CS: 90.04
	[45]	2020	红外+3D骨骼	CS: 91.8, CV: 94.9
	[15]	2020	RGB+2D骨骼	CS: 91.99
	[14]	2020	RGB+3D骨骼	CS: 95.5, CV: 98.0
	[29]	2020	RGB+2D骨骼	CS: 95.45, CV: 98.59
	[43]	2021	RGB+2D骨骼	CS: 91.7
	[23]	2022	RGB+2D骨骼	CS: 96.2

综上所述, 随着对早期融合方法的研究深入, 早期融合的识别准确率越来越高, 相比晚期融合有着更好的效果, 早期融合或将成为多模态融合未来主要的研究方向.

6 总结和展望

本文总结了多模态晚期融合和早期融合方法, 将晚期融合常用方法分为预测分数取平均值、预测分数取最大值、预测分数相乘、预测分数相加、注意力机制、知识蒸馏等方式, 其中使用注意力机制进行晚期融合的方式对性能的提升最为明显, 但由于晚期融合忽略了不同模态特征之间的相关性阻碍了性能的进一步提升; 另外, 本文将多模态早期融合的常用方法分为特征图相加、特征图堆叠、卷积、多模态融合结构搜索、注意力机制等方法, 其中多模态融合结构搜索以及注意力机制的方法对性能的提升最为明显. 本文还对晚期融合和早期融合提升的性能进行了比较, 可以看出随着越来越有效的早期融合方法的提出, 早期融

合可以更有效地利用网络提取的不同模态的不同层次的语义信息, 带来的性能提升已经超过了晚期融合, 通过这些分析和比较, 可以归纳出早期融合或将成为多模态融合未来主要的研究方向.

参考文献

1 Zhao L, Wang SQ, Wang SS, *et al*. Enhanced surveillance video compression with dual reference frames generation. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1592–1606. [doi: 10.1109/TCSVT.2021.3073114]

2 Lu MQ, Hu YC, Lu XB. Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Applied Intelligence, 2020, 50(4): 1100–1111. [doi: 10.1007/s10489-019-01603-4]

3 Xu P, Liu K, Xiang T, *et al*. Fine-grained instance-level sketch-based video retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(5): 1995–2007. [doi: 10.1109/TCSVT.2020.3014491]

4 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. [doi: 10.15888/j.cnki.csa.007875]

5 Ren ZL, Zhang QS, Gao XY, *et al*. Multi-modality learning for human action recognition. Multimedia Tools and Applications, 2021, 80(11): 16185–16203. [doi: 10.1007/s11042-019-08576-z]

6 Vielzeuf V, Lechervy A, Pateux S, *et al*. Centralnet: A multilayer approach for multimodal fusion. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 575–589.

7 Sun ZH, Ke QH, Rahmani H, *et al*. Human action recognition from various data modalities: A review. arXiv:2012.11866, 2020.

8 Liu XL, Li ZB. Deeply fusing multi-model quality-aware features for sophisticated human activity understanding. Signal Processing: Image Communication, 2020, 84: 115809. [doi: 10.1016/j.image.2020.115809]

9 Gavrilyuk K, Sanford R, Javan M, *et al*. Actor-transformers for group activity recognition. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 836–845.

10 Yan A, Wang YL, Li ZF, *et al*. PA3D: Pose-action 3D machine for video recognition. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7914–7923.

11 Zhao JJ, Snoek CGM. Dance with flow: Two-in-one stream action detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9927–9936.

12 Lipton AJ, Fujiyoshi H, Patil RS. Moving target classification and tracking from real-time video. Proceedings of the 4th IEEE Workshop on Applications of Computer Vision. Princeton: IEEE, 1998. 8–14.

13 Pérez-Rúa JM, Vielzeuf V, Pateux S, *et al*. MFAS: Multimodal fusion architecture search. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6959–6968.

14 Das S, Sharma S, Dai R, *et al*. VPN: Learning video-pose embedding for activities of daily living. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 72–90.

15 Joze HRV, Shaban A, Iuzzolino ML, *et al*. MMTM: Multimodal transfer module for CNN fusion. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13286–13296.

16 Xu WY, Wu MQ, Zhao M, *et al*. Fusion of skeleton and RGB features for RGB-D human action recognition. IEEE Sensors Journal, 2021, 21(17): 19157–19164. [doi: 10.1109/JSEN.2021.3089705]

17 Munro J, Damen D. Multi-modal domain adaptation for fine-grained action recognition. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 119–129.

18 Wu HB, Ma X, Li YB. Spatiotemporal multimodal learning with 3D CNNs for video action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 202, 32(3): 1250–1261.

19 贾紫婷. 多模态医学图像融合的关键技术研究. 计算机时代, 2020, (7): 4–6, 11.

20 Li YX, Lu ZC, Xiong XH, *et al*. PERF-Net: Pose empowered RGB-flow net. Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022. 798–807.

21 Xu C, Wu X, Li YC, *et al*. Cross-modality online distillation for multi-view action recognition. Neurocomputing, 2021, 456: 384–393. [doi: 10.1016/j.neucom.2021.05.077]

22 Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423–443. [doi: 10.1109/TPAMI.2018.2798607]

23 Duan HD, Zhao Y, Chen K, *et al*. Revisiting skeleton-based action recognition. arXiv:2104.13586, 2021.

24 Abavisani M, Joze HRV, Patel VM. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1165–1174.

25 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1725–1732.

26 Crasto N, Weinzaepfel P, Alahari K, *et al*. MARS: Motion-augmented RGB stream for action recognition. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7874–7883.

27 Qiu ZF, Yao T, Ngo CW, *et al*. Learning spatio-temporal representation with local and global diffusion. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12048–12057.

28 Liang XY, Qian YH, Guo Q, *et al*. AF: An association-based fusion method for multi-modal classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. [doi: 10.1109/TPAMI.2021.3125995]

29 Davoodikakhki M, Yin KK. Hierarchical action classification with network pruning. Proceedings of the 15th International Symposium on Visual Computing. San Diego: Springer, 2020. 291–305.

30 Bian CL, Feng W, Wan L, *et al*. Structural knowledge distillation for efficient skeleton-based action recognition. IEEE Transactions on Image Processing, 2021, 30: 2963–2976. [doi: 10.1109/TIP.2021.3056895]

31 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. Las Vegas: IEEE, 2016. 1010–1019.

32 Liu J, Shahroudy A, Perez M, *et al*. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2684–2701.

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

34 Kuehne H, Jhuang H, Garrote E, *et al*. HMDB: A large video database for human motion recognition. Proceedings of 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 2556–2563.

35 Li J, Liu XL, Zhang WX, *et al*. Spatio-temporal attention networks for action recognition and detection. IEEE Transactions on Multimedia, 2020, 22(11): 2990–3001. [doi: 10.1109/TMM.2020.2965434]

36 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1302–1310.

37 Kreiss S, Bertoni L, Alahi A. PifPaf: Composite fields for human pose estimation. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11969–11978.

38 Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 2938–2946.

39 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5696.

40 Das S, Dai R, Koperski M, *et al*. Toyota smarthome: Real-world activities of daily living. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 833–842.

41 Rogez G, Weinzaepfel P, Schmid C. LCR-Net++: Multi-person 2D and 3D pose detection in natural images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(5): 1146–1161.

42 Stroud JC, Ross DA, Sun C, *et al*. D3D: Distilled 3D networks for video action recognition. Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020. 614–623.

43 Moon G, Kwon H, Lee KM, *et al*. Integralaction: Pose-driven feature integration for robust human action recognition in videos. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021. 3334–3343.

44 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. Salt Lake City: IEEE, 2018. 1159–1168.

45 De Boissiere AM, Noumeir R. Infrared and 3D skeleton feature fusion for RGB-D action recognition. IEEE Access, 2020, 8: 168297–168308. [doi: 10.1109/ACCESS.2020.3023599]

(校对责编: 孙君艳)