软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(Suppl.(2)):1−15

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

关联模态补偿的视频动作识别算法*

宋思捷, 刘家瑛, 厉扬豪, 郭宗明

(北京大学 计算机科学技术研究所,北京 100871)

通讯作者: 郭宗明, E-mail: guozongming@pku.edu.cn

摘 要: 随着深度摄像机的发展,不同模态的视频数据更易获得.基于多模态数据的视频动作识别也受到越来越广泛的关注.不同模态的数据能够从多个角度对视频动作进行描述,如何有效地利用多模态数据并形成优势互补是视频动作识别中的重要方向.提出了一种基于关联模态补偿的视频动作识别算法.该方法以RGB和光流场视频数据为源模态,以3D骨架数据为辅助模态,利用源模态和辅助模态高层特征空间的关联性,补偿源模态的特征提取.该算法基于卷积神经网络和长短期记忆网络,对源模态数据和辅助模态数据进行时空域特征建模.在此基础上,提出了基于残差子网络的模态适应模块,通过统一源模态特征和辅助模态特征的数据分布,实现辅助模态对源模态的特征补偿.考虑到源模态数据和辅助模态数据在动作类别或动作样本等方面存在不同程度的对齐情况,设计了多层次模态适应算法,以适应不同的训练数据.所提算法仅在训练过程中需要辅助模态的帮助,在测试过程中可以仅根据源模态数据进行动作的识别,极大地拓展了该算法的实用性.在通用公共数据集上的实验结果表明,相比于现有动作识别算法,该算法取得了更加优越的性能.

关键词: 视频动作识别;多模态数据;关联模态补偿;深度学习;残差学习

中文引用格式: 宋思捷,刘家瑛,厉扬豪,郭宗明.关联模态补偿的视频动作识别算法.软件学报,2018,29(Suppl.(2)):1−15. http://www.jos.org.cn/1000-9825/18013.htm

英文引用格式: Song SJ, Liu JY, Li YH, Guo ZM. Modality compensation based action recognition. Ruan Jian Xue Bao/Journal of Software, 2018,29(Suppl.(2)):1−15 (in Chinese). http://www.jos.org.cn/1000-9825/18013.htm

Modality Compensation Based Action Recognition

SONG Si-Jie, LIU Jia-Ying, LI Yang-Hao, GUO Zong-Ming

(Institute of Computer Science and Technology, Peking University, Beijing 100871, China)

Abstract: With the prevalence of depth cameras, video data of different modalities become more common. Multi-Modal data based human action recognition attracts increasing attention. Different modal data describe human actions from distinct perspectives. How to effectively utilize the complementary information of multi-modal data is a key topic in this area. In this study, we propose a modality compensation based method for action recognition. With RGB/optical flow as source modal data and skeletons as auxiliary modal data, we aim to compensate the feature learning from source modal data, through exploring the common spaces between source and auxiliary modalities. The proposed model is based on deep convolutional neural network (CNN) and long short term memory (LSTM) network to extract spatial and temporal features. With the help of residual learning, a modality adaptation block is proposed to align the distributions of different modalities and achieve modality compensation. To deal with different alignment of source and auxiliary modal data, we propose hierarchical modality adaptation schemes. The proposed model only requires auxiliary modal data in the training process, and is able to improve the recognition performance only with source modal data in the testing phase, which expands the application scenarios of the proposed model. The experiment results illustrate that proposed method outperforms other state-of-the-art approaches.

Key words: action recognition; multi-modal data; modality compensation; deep learning; residual learning

* 基金项目: 国家自然科学基金(61772043)
Foundation item: National Natural Science Foundation of China (61772043)
收稿时间: 2018-04-13; 修改时间: 2018-06-13; 采用时间: 2018-09-30

视频动作识别在视频分析与理解领域中是一项基础而重要的课题,其在视频监控、人机交互、虚拟现实等方面有广泛的应用前景和巨大的市场需求,受到了学术界和工业界的广泛关注.然而,由于多变的拍摄角度和尺度、多样的光照条件和场景背景以及视频动作的复杂性,视频动作识别仍然是一个极具挑战性的研究课题.

多数视频动作识别算法以 RGB 视频数据为研究对象[1].早期的算法依赖于手动提取视频特征[2],随着深度学习的蓬勃发展,基于神经网络的视频动作识别算法在视频特征提取方面展现了强大的建模能力,并取得了良好的实验性能.其中,Two-Stream 模型是一套经典的动作识别框架[3],包括空域网络和时域网络两个支路.空域支路对视频帧内空间特征建模,时域支路则描述了视频帧间时域特征,最后通过融合两个支路的概率分布,得到最终的动作识别结果.后续很多工作围绕 Two-Stream 模型的基本框架展开,在时空域特征学习或时空域支路的关联性等方面进行了改进[4−7].然而,在基于 Two-Stream 模型的视频动作识别算法中,仍存在以下问题.

- 在训练样本数量有限的情况下,难以提取视频的不变性特征,并覆盖所有的动作模式(如同一动作在拍摄角度和尺度上的差异性).
- 对于空域支路,空域特征易受到动作背景的干扰,导致动作前景和背景的混合,降低了特征的判别力.
- 对于时域支路,对光流场数据的卷积操作易在特征中引入无关背景动作噪声,破坏特征的鲁棒性.

为了解决上述问题,本文引入 3D 骨架数据,从多模态数据的互补性出发,挖掘不同数据模态间的关联关系,增强 Two-Stream 模型中视频数据特征提取的判别力和鲁棒性,从而提升动作识别的准确度.随着深度相机(如 Microsoft Kinect)的发展[8]和人体姿态估计技术[9]的成熟,3D 骨架数据较以往更加普遍.作为人体的高级特征表示,3D 骨架数据对光照条件较为鲁棒,具有视角不变性和尺度不变性.同时,3D 骨架不受背景干扰,能够提供清晰的人体运动特征和姿态信息.因此,基于 3D 骨架的视频动作识别在近年来吸引了越来越多的关注[10−13].然而,在仅依赖 3D 骨架的动作识别中,外观信息的缺失易对动作识别过程引起歧义,影响动作识别的准确度.

为了弥补 3D 骨架数据缺少外观信息这一缺点,一些工作将 3D 骨架数据和其他模态数据相结合,利用关联模态的互补信息来加以弥补.有些工作借鉴 Two-Stream 的思路,注重在不同模态支路融合的方面,如 Zolfaghari 等人[14]利用马尔可夫链,将外观信息、运动信息和人体姿态结合以完成视频动作识别.其他工作则利用跨模态特征学习的方法,探索不同模态数据的高层特征空间的相关关系,实现特征互补[15,16].Mahasseni 等人[16]提出,基于 3D 骨架的特征表示对 RGB 数据特征是一种有效的补充,即通过 3D 骨架的特征空间约束 RGB 数据的特征提取,有效挖掘 RGB 数据中的固有特征.然而,在这项工作中,没有考虑 3D 骨架数据和光流场数据的相互结合,限制了动作识别的精度提升.因此,本文提出一种通用的方法,考虑 3D 骨架与 RGB 数据或光流场数据的模态关联性,将 3D 骨架数据同时适应于 RGB 数据和光流场数据.此外,本文考虑了在训练时多模态数据的对齐情况,如 3D 骨架和 RGB 视频不存在对应关系,或 3D 骨架和 RGB 视频一一对应.

本文将 RGB 和光流场数据作为源模态,将 3D 骨架数据作为辅助模态,提出一种基于关联模态补偿的视频动作识别算法.本文以 Two-Stream 模型为基础,使用卷积神经网络(convolutional neural network,简称 CNN)和长短期记忆神经网络(long short term memory,简称 LSTM)构建算法框架,主要思想是通过适应性特征学习,使源模态数据特征能够表达辅助模态的优点,以补偿源模态表征.在空域网络中,利用 3D 骨架数据对背景的鲁棒性以及对运动信息的刻画,鼓励网络表达视频前景和外观细节,并在一定程度上考虑帧间动态关系.对于时域网络,3D 骨架则主要用来补充姿态信息,并且引导网络在提取时域特征时规避背景的运动噪声.本文算法仅在训练过程中需要辅助模态数据,在测试中可仅依赖源模态数据做出动作识别.本文贡献主要表现在以下 3 个方面.

(1) 本文提出了基于关联模态补偿的视频动作识别算法,通过辅助模态的适应特征学习,补充源模态数据特征.

(2) 本文提出了多层次模态适应方法,包括模态层次、类别层次和样本层次,应用于不同对齐方式的训练数据,充分利用关联模态数据的互补关系.

(3) 视频动作识别的实验结果表明,本文算法能够有效增强视频时空域特征的鲁棒性和判别力,在通用公共数据集上的识别准确度优于其他现有算法.

本文第 1 节对相关工作进行简要介绍,主要包括视频动作识别和多模态特征学习两个方面.第 2 节介绍本

文提出的基于关联模态补偿的视频动作识别算法,借助辅助模态数据的优势,通过模态适应模块对源模态数据进行特征补偿.第 3 节通过实验将本文算法与其他现有视频动作识别算法进行比较,以证明本文算法的有效性.最后对全文进行总结.

1 相关工作

1.1 视频动作识别

根据视频动作识别的研究对象,视频动作识别可划分为基于 RGB 视频的动作识别算法和基于 3D 骨架的动作识别算法.在视频图像领域的研究问题中,如何提取鲁棒性较好的特征是研究人员致力研究的问题[48,49].在基于 RGB 视频的动作识别算法中,传统算法[2,17]在视频的特征表示方面主要依靠手动特征描述子,如梯度直方图(histogram of oriented gradient,简称 HOG)、光流直方图(histogram of flow,简称 HOF)和运动边界直方图(motion boundary histogram,简称 MBH)等等,然后运用特征袋(bag of feature,简称 BOF)等方式对兴趣点进行编码,最后使用分类器,如支持向量机(support vector machine,简称 SVM)等对编码后的特征进行分类.随着大规模视频数据集的出现,传统的视频特征描述子难以满足视频表征的需求,而基于深度神经网络的算法能够自动地根据视频数据完成时空域建模,同时做出动作类别的判断.Two-Stream 模型是使用卷积神经网络进行动作识别的早期工作[3],也成为了很多后续工作的基础框架[4−7].Two-Stream 模型对空域和时域分别建模,一定程度上丢失了视频空域和时域的内在联系.为了能够同时对视频时空域进行刻画,Tran 等人[20]提出了 3D 卷积神经网络(C3D)的概念,运用 3D 卷积核,扩展视频特征提取的维度.在 C3D 的基础上,VLAD3 算法[21]能够对视频进行多层次建模,表达短期、中期和长期的时域动态特征.此外,Fernando 等人[22,23]提出了基于排序池化(rank pooling)的视频动作识别算法,对视频序列的动态信息进行编码,并生成对应的视频动态图,有效提升了捕捉视频关键信息的能力.为满足实际应用中的性能需求,Wang 等人[6]提出了时域分块网络(temporal segment network,简称 TSN),将视频动作识别的准确度提升到新的高度.此外,循环神经网络(recurrent neural network,简称 RNN)被广泛应用在序列建模的任务上,也同样适用于视频数据的建模,实现时域特征的自动学习[24−26],此类算法一般先利用卷积神经网络对每帧视频进行表征,然后运用循环神经网络对时域动态建模.

视频动作识别的另一个重要分支为基于 3D 骨架的视频动作识别算法.早期的工作应用生成式模型[27,28],如隐马尔可夫模型(hidden Markov model,简称 HMM),用来描述时域动态信息.然而,这些模型的参数较多,难以有效地在参数空间中寻找最优解.还有一些基于 3D 骨架的动作识别算法遵从特征提取-特征编码-特征分类的传统思路[29,30].由于骨架数据的三维坐标表示比较简洁,易以旋转和平移等几何变换描述几何关系特征,或者以速度、加速度等描述动力学特征.然后将这些局部特征进行编码后输入分类器中,得到最终的识别结果[31].随着深度学习的发展,循环神经网络也逐渐被应用于基于 3D 骨架的视频动作识别中,自动地从三维坐标中提取时空域特征[10−13].最近的工作将三维骨架序列转化为二维骨架图,从而将对序列的分类问题转化为对图像的分类问题,并利用卷积神经网络对其进行分类,有效提升了动作识别的准确度[32].为了弥补 3D 骨架数据没有外观信息缺失的缺点,围绕多模态数据的研究逐渐展开.Zolfaghari 等人[14]利用马尔可夫链,在概率模型的约束下,将外观信息、运动信息和人体姿态有机结合.其他工作则利用跨模态特征学习的方法,探索不同模态数据的高层特征空间[15,16].Shi 等人[15]提出利用循环神经网络从深度图和 3D 骨架的相关关系中学习特权信息(priviledged information,简称 PI).Mahasseni 等人[16]提出利用 3D 骨架数据特征空间约束 RGB 的特征空间,在基于 RGB 的动作识别上取得了更好的效果.

1.2 多模态特征学习

不同模态的特征能够提供互补信息,多模态特征学习近来受到广泛关注和研究,旨在通过探索不同模态的公共空间或者特异空间,优化特征提取过程.一种多模态特征学习的思路通过约束不同模态之间的数据表征对齐各模态的数据分布[33,34],本质上可视为领域自适应(domain adaptation)问题[35].Long 等人引入了深度领域适应网络,最小化源域和目标域高层特征空间的最大平均差异(maximum mean discrepancy,简称 MMD).然

而,Bousmalis 等人[36]指出,只关注源域和目标域的公共特征空间会忽略不同模态的特异性.鉴于此,Wang 等人[33]提出了一个多模态特征学习统一框架,能够同时对不同模态的公共空间和特异空间进行建模.另有工作对公共空间和特异空间进行显式约束[34],在最小化多模态数据特征公共空间的同时,考虑最大化多模态特异空间的距离.此外,多模态数据之间的关系可以直接由网络建模,Schrode 等人[50]利用全卷积网络自动刻画 RGB 和深度图像的联系,建立像素级别的概率预测.Xu 等人[51]利用 RGB 图像恢复热力图像,利用中间特征进行行人检测.Kiela 等人[52]在文本数据和图像数据之间建立双线性门模型,有效提升了大规模数据分类的性能.

2 基于关联模态补偿的视频动作识别算法

2.1 基础结构

如图 1 所示,本文算法以 Two-Stream 模型为基本框架.空域支路从 RGB 视频帧中提取外观信息,时域支路从堆叠的水平和竖直方向的光流场中提取动态运动信息.每个支路由卷积神经网络和长短期记忆神经网络组成.其中,卷积神经网络用来编码输入的空域信息,长短期记忆神经网络收集并整理时间窗口内接收到的信息.本文算法中,空域支路和时域支路分别独立训练,最终的动作识别结果通过融合两个支路的概率分布得到.图 1 用上标 r 和 f 分别表示处理 RGB 输入的空域支路和处理光流场输入的时域支路.下文对本文算法的描述仅针对单个支路,为了简化符号表示,以下将上标 r 和 f 省略.本文以 Two-Stream 模型为基础,设计了基于残差子网络的模态适应模块,对源模态的数据特征做出补偿.

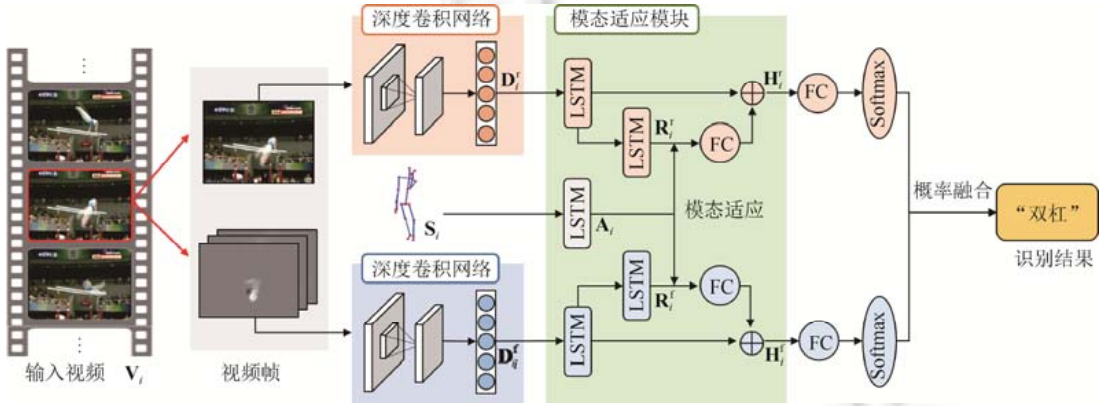


Fig.1 The framework of modality compensation based action recognition

图 1 基于关联模态补偿的视频动作识别算法框架

卷积神经网络:与传统的手动提取视频特征相比,卷积神经网络在特征提取方面展现了极其强大的建模能力[18,19].给定当前批量数据(batch)中第 i 个视频序列 $\mathbf{V}_i=\{\mathbf{v}_{i,t}:t=1,\ldots,T\}$,其中,$\mathbf{v}_{i,t}$ 为时刻 t 的视频帧,本文算法使用卷积神经网络对其进行空间特征提取.受文献[6]的启发,在实验中以 BN-Inception 网络[37]的 global-pool 层的输出为特征,每帧视频可由维度为 1 024 的向量 $\mathbf{d}\in\mathbb{R}^{1024}$ 表示.接下来,抽取的卷积特征序列 $\mathbf{D}_i=\{\mathbf{d}_{i,t}:t=1,\ldots,T\}$ 将输入到长短期记忆神经网络.

长短期记忆神经网络:本文使用长短期记忆神经网络[38]探索视频序列的时域关联性.循环神经网络的自连接结构能够处理和保存历史信息,可有效利用长期或短期特征,但是存在梯度消失或梯度爆炸等缺点.长短期记忆神经网络通过输入单元、记忆单元和输出单元的控制,解决网络优化中的距离依赖问题,避免梯度消失或梯度爆炸情况的出现.

2.2 关联模态适应特征学习

在视频分类问题中,同一种类内部的多样性和不同种类之间的相似性导致难以提取到鲁棒性较好的特征,以适应不同的拍摄视角和尺度.此外,杂乱的背景容易在特征提取过程中引入噪声.为了解决上述问题,本文提

出了一种关联模态适应特征学习的方法,补充输入源视频的特征表示,使其更具鲁棒性.而 3D 骨架数据作为一种人体的高级特征表示,具有视角不变性和尺度不变性,同时能够排除视频背景的干扰.本文旨在利用 3D 骨架数据的这些优点改善源模态数据的特征学习.

模态适应模块:本文使用嵌入各支路的模态适应模块完成关联模态适应特征学习.对于每个支路,模态适应模块包括主网络和残差子网络[19],如图 1 所示.其中,主网络和子网络均由循环层组成.一方面,主网络能够维持源模态数据中的原始信息.另一方面,残差子网络的旁路能够通过适应学习对原始特征进行补偿.残差旁路的基本定义为

$$\mathbf{z}^{l+1}=F(\mathbf{z}^{l},\{\theta^{l}\})+\mathbf{z}^{l} \tag{1}$$

其中,$\mathbf{z}^{l}$ 和 $\mathbf{z}^{l+1}$ 分别为当前层的输入和输出,F 为非线性残差映射,θ^{l} 为映射函数中的参数.如图 1 所示,本文算法采用一层循环层和一层全连接层(fully connected,简称 FC)作为残差映射函数.

本文算法适应特征学习的原理在于将源模态特征空间变换到辅助模态的特征空间,使得源模态特征能够表达辅助模态的优势.对于一个给定的 3D 骨架序列 $\mathbf{S}_i=\{\mathbf{s}_{i,t}:t=1,\ldots,T\}$,首先使用经过预训练的长短期记忆网络将其编码为特征向量 $\mathbf{A}_i=\{\mathbf{a}_{i,t}:t=1,\ldots,T\}$.在模态适应模块中,将第 2 个循环层的输出作为源模态的特征表示,如图 1 所示,记为 $\mathbf{R}_i=\{\mathbf{r}_{i,t}:t=1,\ldots,T\}$.为了完成空间转换,本文通过优化源模态数据的特征学习,使得源模态和辅助模态特征空间相似,可由最小化下式实现:

$$dist(\overline{\mathbf{R}},\overline{\mathbf{A}}) \tag{2}$$

其中,$\overline{\mathbf{R}}$ 和 $\overline{\mathbf{A}}$ 分别为源模态和辅助模态特征空间.然而,由于三维骨架和彩色视频之间存在数据“鸿沟”,难以直接定义函数 $dist(\cdot)$.本文通过对齐源模态和辅助模态数据分布的方式达到空间转换的目的,详见下一小节.

2.3 多层次模态适应

在关联模态适应过程中,根据训练数据的特点,需要考虑以下两种场景.

- 通用场景:训练数据中 RGB 视频和 3D 骨架视频不存在对应关系.
- 特定场景:训练数据中 RGB 视频和 3D 骨架视频一一对应.

为了适应不同的训练场景,以充分利用源模态数据和辅助模态数据的互补关系,本文设计了不同层次的关联模态适应方法,包括模态层次、类别层次和样本层次.表 1 总结了各个模态适应方法对训练数据的要求.

Table 1 Requirements for multi-modal data of each modality adaptation scheme

表 1 各模态适应方法对训练数据的要求

适应层次	数据对齐要求	适用场景
模态层次	无对齐要求	通用场景/特定场景
类别层次	RGB 和 3D 视频类别对齐	特定场景
样本层次	RGB 和 3D 视频一一对应	特定场景

2.3.1 模态层次

模态层次的模态适应对训练数据中源模态和辅助模态没有任何对齐要求,是适用范围最广泛的一种关联模态适应方法.受到领域自适应的启发,本文采用最小化最大平均差异(MMD)的方式对齐源模态和辅助模态的数据分布.给定来自两个不同分布的样本集,$X=\{\mathbf{x}_i\}_{i=1}^{m_x}$ 和 $Y=\{\mathbf{y}_i\}_{i=1}^{m_y}$,最大平均差异可由下式计算:

$$\left.\begin{aligned}\mathrm{MMD}^2[X,Y]&=\left\|E_x\left[\phi(\mathbf{x})\right]-E_y\left[\phi(\mathbf{y})\right]\right\|^2\\&=\frac{1}{m_x^2}\sum_{i=1}^{m_x}\sum_{i'=1}^{m_x}\phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_{i'})+\frac{1}{m_y^2}\sum_{j=1}^{m_y}\sum_{j'=1}^{m_y}\phi(\mathbf{y}_j)^{\mathrm{T}}\phi(\mathbf{y}_{j'})\\&\quad-\frac{2}{m_x m_y}\sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{y}_j)\end{aligned}\right\} \tag{3}$$

其中,ϕ为特征映射函数,m_x 和 m_y 分别为两个样本集的样本数量.

对处于不同分布下的源模态数据空间和辅助模态数据空间,本文算法以视频粒度的特征表示作为空间中

的特征向量,并以此为样本计算两个分布下的最大平均差异.对于一组批量数据,包含 n 个源模态数据样本和 n 个辅助模态特征样本.对于当前批量中第 i 个源模态视频 $\mathbf{V}_i$,其编码后的特征为 $\mathbf{R}_i$,则其视频粒度的特征表示为

$$\hat{\mathbf{r}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{i,t} \tag{4}$$

同理,若第 i 个辅助模态数据 $\mathbf{S}_i$ 编码后的特征为 $\mathbf{A}_i$,则其视频粒度的特征表示为

$$\hat{\mathbf{a}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{a}_{i,t} \tag{5}$$

令 $k(\mathbf{x},\mathbf{y})=\langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$,则源模态和辅助模态的最大平均差异为

$$d_D = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}k(\hat{\mathbf{a}}_i,\hat{\mathbf{a}}_{i'}) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{j'=1}^{n}k(\hat{\mathbf{r}}_j,\hat{\mathbf{r}}_{j'}) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(\hat{\mathbf{a}}_i,\hat{\mathbf{r}}_j) \tag{6}$$

其中,n 为批量数据中的样本数量(batch size),线性核函数 $k(\mathbf{x},\mathbf{y})=\mathbf{x}^{\mathrm{T}}\mathbf{y}$.

2.3.2 类别层次

样本的类别信息可以为特征提取带来更加明确的指导.类别层次的关联模态适应旨在使得同一类别下源模态和辅助模态的特征空间更加相似,以简化空间变换的学习.在类别层次的关联模态适应中,本文在计算最大平均差异时将样本的类别信息考虑在内.根据视频粒度的样本特征,基于类别的最大平均差异定义为

$$d_C = \frac{1}{\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{i,i'}}\left(\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{i,i'}\times k(\hat{\mathbf{a}}_i,\hat{\mathbf{a}}_{i'}) + \sum_{j=1}^{n}\sum_{j'=1}^{n}u_{j,j'}\times k(\hat{\mathbf{r}}_j,\hat{\mathbf{r}}_{j'}) - 2\sum_{i=1}^{n}\sum_{j=1}^{n}u_{i,j}\times k(\hat{\mathbf{a}}_i,\hat{\mathbf{r}}_j)\right) \tag{7}$$

其中,批量数据中包含 n 个源模态样本和 n 个辅助模态样本,若批量数据中第 i 个源模态样本和第 j 个辅助模态样本属于同一类别,则 $u_{i,j}=1$,否则为 0.

2.3.3 样本层次

在样本层次的关联模态适应中,需要满足源模态视频数据和3D骨架数据存在一一对应的关系,是粒度最精细的模态适应.一般来讲,可以直接利用源模态和辅助模态的帧间对应关系.然而,虽然输入源模态视频的每一帧都与一帧 3D 骨架数据对应,但在动作幅度比较小的情况下,光流场数据难以传达有效信息.因此,对于光流场数据来说,最小化对应的帧间关系并不合理.另一方面,实验发现,对于 RGB 数据,由于 3D 骨架数据中存在噪声,最小化帧间距离并没有带来性能提升.因此,在样本层次的模态适应中,本文算法仍使用视频粒度的特征表示,并采用欧氏距离(Euclidean distance)约束适应特征的学习,可形式化表示如下:

$$d_S = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{a}}_i - \hat{\mathbf{r}}_i\right\|^2 \tag{8}$$

其中,$\hat{\mathbf{a}}_i$ 和 $\hat{\mathbf{r}}_i$ 分别为第 i 个视频样本辅助模态和源模态的特征描述子,n 为批量数据中的样本个数.

2.4 损失函数

为了获得对视频序列 $\mathbf{V}_i$ 的识别结果,本文算法将模态适应模块输出的隐层状态 $\mathbf{H}_i=[\mathbf{h}_{i,1},\ldots,\mathbf{h}_{i,T}]$进行整合,并映射到维度为 C 的向量 $\mathbf{g}_i=[g_{i,1},\ldots,g_{i,C}]$,以表示 C 类动作的置信度,即:

$$\mathbf{g}_i = \mathbf{W}_g\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_{i,t}\right) + \mathbf{b}_g \tag{9}$$

其中,$\mathbf{W}_g$ 和 $\mathbf{b}_g$ 为 softmax 层前最后一个全连接层的参数.视频 $\mathbf{V}_i$ 属于第 c 类的概率为

$$p(g_{i,c}\mid\mathbf{V}_i) = e^{g_{i,c}}\Big/\sum_{j=1}^{C}e^{g_{i,j}}\ (c=1,\ldots,C) \tag{10}$$

本文算法每一个支路的优化目标包括两部分,即分类的交叉熵损失以及源模态和辅助模态特征空间的最大平均差异,其形式化表示为

$$L = -\sum_{i=1}^{n}\sum_{c=1}^{C}l_{i,c}\log p(g_{i,c}) + \lambda_d d \tag{11}$$

其中,若第 i 个视频的类别为 c,则 $l_{i,c}=1$,否则为 0,常数 λ_d 用来平衡损失函数中每一项的权重,根据不同训练数据的对齐情况,以最大平均差异表示的关联模态适应损失 $d\in\{d_D,d_C,d_S\}$.

3 实验结果

为验证本文算法的有效性,本文在 NTU RGB+D 数据集[12]和 UCF-101 数据集[39]上开展实验.实验部分中详细探讨了算法每一部分的作用.注意,本文算法仅在训练过程中需要辅助模态数据,在测试阶段可仅依赖源模态数据.

3.1 实验设定

NTU RGB+D 数据集[12]是迄今为止最大的视频动作识别数据集,包括对齐的多模态数据(RGB,3D 骨架,深度图,红外图).该数据集包含 56 880 个视频样本,共计 400 多万帧,提供了 60 类动作,包括单人动作和双人交互动作.为全面评估算法性能,该数据集提供了两种评测方式:交叉视角(cross view,简称 CV)和交叉人物(cross subject,简称 CS).由于该数据集中 RGB 视频每帧的分辨率大小为 1920×1080,为适应网络输入,需将每帧视频分辨率大小调整为 224×224,导致人物只在画面中占据一小部分.为避免不必要的性能损失,本文在实验中将每帧 RGB 视频进行裁剪.首先将 3D 骨架数据映射到 RGB 视频帧中,可得人物在画面中占据坐标的范围(x_{min}:x_{max}, y_{min}:y_{max}),然后裁取当前帧范围为[x_{min}−250:x_{max}+250,y_{min}−50:y_{max}+50]的区域,以增加人物的分辨率,并在裁剪后的视频上提取光流场数据[6].为了加速光流场数据的提取过程,对 RGB 和 3D 骨架数据在时域上进行步长为 5 的下采样.此外,本文在实验中对 3D 骨架进行归一化处理[12]以实现视角不变性和位置不变性.

UCF-101 数据集[39]包含 13 000 个视频样本,提供了 101 个动作类别.每个视频样本时长约为 3s~10s,包含 100~300 帧,每帧分辨率为 320×240.该数据集因其动作类别的多样性和背景的复杂性极具挑战性.本文采用文献[6]中的评测方法,报告算法在 3 种训练集/测试集划分方法下的平均性能.该数据集中的所有视频统一时域下采样为 60 帧.由于 UCF-101 数据集不提供 3D 骨架数据,实验中将 NTU RGB+D 数据集的 3D 骨架数据作为辅助模态.

实验设定细节:对于辅助模态数据,使用经过预训练的长短期记忆神经网络对 3D 骨架数据编码,其中,该长短期记忆神经网络由一层循环层组成,包含 1 024 个神经元.对于源模态数据,本文实验在两个数据集上分别对 BN-Inception 网络[6]进行预训练,并将 global-pool 层的输出作为卷积特征,输入接下来的长短期记忆神经网络.本文采用包含两层循环层的长短期记忆神经网络,每层包含 1 024 个神经元,其中一层循环层在残差旁路中,如图 2(a)所示.全连接层中的神经元个数同样设置为 1 024.在实际的网络训练过程中,用于编码 3D 骨架的长短期记忆网络中的参数被固定下来,只训练用于处理 RGB 的空域支路和用于处理光流场的时域支路.由于 GPU 的显存限制,NTU RGB+D 数据集中批量数据的大小设为 256,UCF-101 数据集中批量数据的大小设为 192.本文采用 Adam 算法[40]作为网络训练的优化器,以自动调整学习率.初始学习率设置为 0.001.此外,采用概率为 0.5 的 Dropout[41]避免过拟合.

(a) Res-LSTM (b) 2-LSTM

Fig.2 The structures of Res-LSTM and 2-LSTM, respectively. The input features are from CNN (omitted in the figure)

图 2 Res-LSTM 和 2-LSTM 结构图,其中,输入来自卷积神经网络(图中省略)

3.2 关联模态适应的有效性

本文在 NTU RGB+D 和 UCF-101 数据集上分别对比了关联模态适应方法,以验证模态适应的有效性.为解决表 1 中展示的不同场景,本文进行了以下实验.

- Res-LSTM:不使用关联模态适应的基础算法,如图 2(a)所示.
- D-Res-LSTM:模态层次的关联模态适应,针对通用场景和特定场景.
- C-Res-LSTM:类别层次的关联模态适应,针对特定场景.

• S-Res-LSTM:样本层次的关联模态适应,针对特定场景.

• J-Res-LSTM:联合模态层次,类别层次和样本层次的关联模态适应,针对特定场景.

表 2 展示了 NTU RGB+D 和 UCF-101 数据集在 RGB 和光流场单个支路下的动作识别性能.NTU RGB+D 数据集提供了一一对应的源模态和辅助模态数据,因此,本文在该数据机上评测了模态层次、类别层次和样本层次的关联模态适应,对应的λ_d分别设置为 100、0.001、1.实验结果表明,关联模态适应能够有效提升动作识别的准确率.D-Res-LSTM 和 C-Res-LSTM 因其关联模态适应的粒度较为粗糙取得了相似的实验结果.S-Res-LSTM 在更细的粒度上将源模态和辅助模态数据分布对齐,并在动作识别的性能上带来了 2%~3%的准确率提升.采用联合模态适应的 J-Res-LSTM 与 S-Res-LSTM 的实验性能相似.主要原因在于 S-Res-LSTM 是粒度最精细的模态适应方法,在进行样本层次的适应后,源模态和辅助模态的特征空间在模态和类别层次上的距离也会缩短.因此,联合使用模态层次、类别层次和样本层次的模态适应(J-Res-LSTM)与 S-Res-LSTM 相比性能相近.

在 UCF-101 数据集中,RGB 视频并没有和 NTU RGB+D 的 3D 骨架对齐,因此,可视为通用场景下的训练数据,对应地,采用模态层次的关联模态适应,并将λ_d设为 100.在辅助模态的帮助下,相比基础算法,模态适应为 UCF-101 上的识别准确度带来了 1%~2%的提升.

Table 2 Performance on the NTU RGB+D and UCF-101 (split 1) datasets in accuracy (%), respectively

表 2 NTU RGB+D 和 UCF-101(split 1)数据集的实验性能(准确率:%)

实验设定	NTU-CS		NTU-CV		UCF-101	
	RGB	光流场	RGB	光流场	RGB	光流场
Res-LSTM	79.6	85.8	85.7	91.8	81.6	83.8
D-Res-LSTM	81.4	86.4	87.3	92.0	**83.3**	**85.7**
C-Res-LSTM	81.5	86.6	87.2	92.1	–	–
S-Res-LSTM	**82.0**	**87.6**	**89.1**	93.3	–	–
J-Res-LSTM	81.8	87.3	**89.1**	**93.8**	–	–

3.3 残差学习的有效性

本文采用残差子网络的设计,以实现关联模态适应,主要出于两方面的考虑:(1) 源模态数据中的原始特征可以通过直连通路保持.(2) 从辅助模态数据中通过模态适应学习到的优势特征可以通过旁路补偿.为验证上述考虑,本文在实验中比较了 Res-LSTM 和普通 LSTM 结构 2-LSTM,如图 2(b)所示.在 2-LSTM 中,每个循环层和全连接层均包含 1 024 个神经元,因此,Res-LSTM 和 2-LSTM 有相同的参数量.

本文在 NTU RGB+D 和 UCF-101 数据集上分别对基于以上两种结构的模态适应进行了实验,实验结果如图 3 所示.对于 2-LSTM,在第 2 个循环层的输出后运用模态适应.本文在 2-LSTM 的基础结构上,对 NTU RGB+D 数据集评测了样本层次的模态适应(S-2-LSTM),对 UCF-101 数据集评测了模态层次的模态适应(D-2-LSTM).在没有应用模态适应时,2-LSTM 和 Res-LSTM 能够实现相似的动作识别的准确度.然而,与模态适应结合后,Res-LSTM 结构显示出了更多的优势,表明残差子网络中的旁路能够有效利用辅助信息,对源模态数据做出补偿.同时,本文提出的模态适应方法同样能够对 2-LSTM 的结构做出提升,进一步说明了模态适应的有效性.

本文以 2-LSTM 为基础,在 NTU RGB+D 数据集上验证了各个模态层次的效果,见表 3.实验结果显示,在大多数情形下,各个层次的模态适应在 2-LSTM 上能够取得提升,进一步验证了本文模态适应的有效性.

Table 3 Performance on the NTU RGB+D dataset with 2-LSTM as backbone in accuracy (%), respectively

表 3 各模态适应层次在 NTU RGB+D 数据集的实验性能(基础结构:2-LSTM,准确率:%)

实验设定	NTU-CS		NTU-CV	
	RGB	光流场	RGB	光流场
2-LSTM	80.6	86.2	85.4	92.8
D-2-LSTM	80.6	86.5	85.6	91.2
C-2-LSTM	80.3	86.6	87.2	91.6
S-2-LSTM	80.6	86.8	87.4	**92.5**
J-2-LSTM	**81.7**	**87.1**	**88.9**	92.3

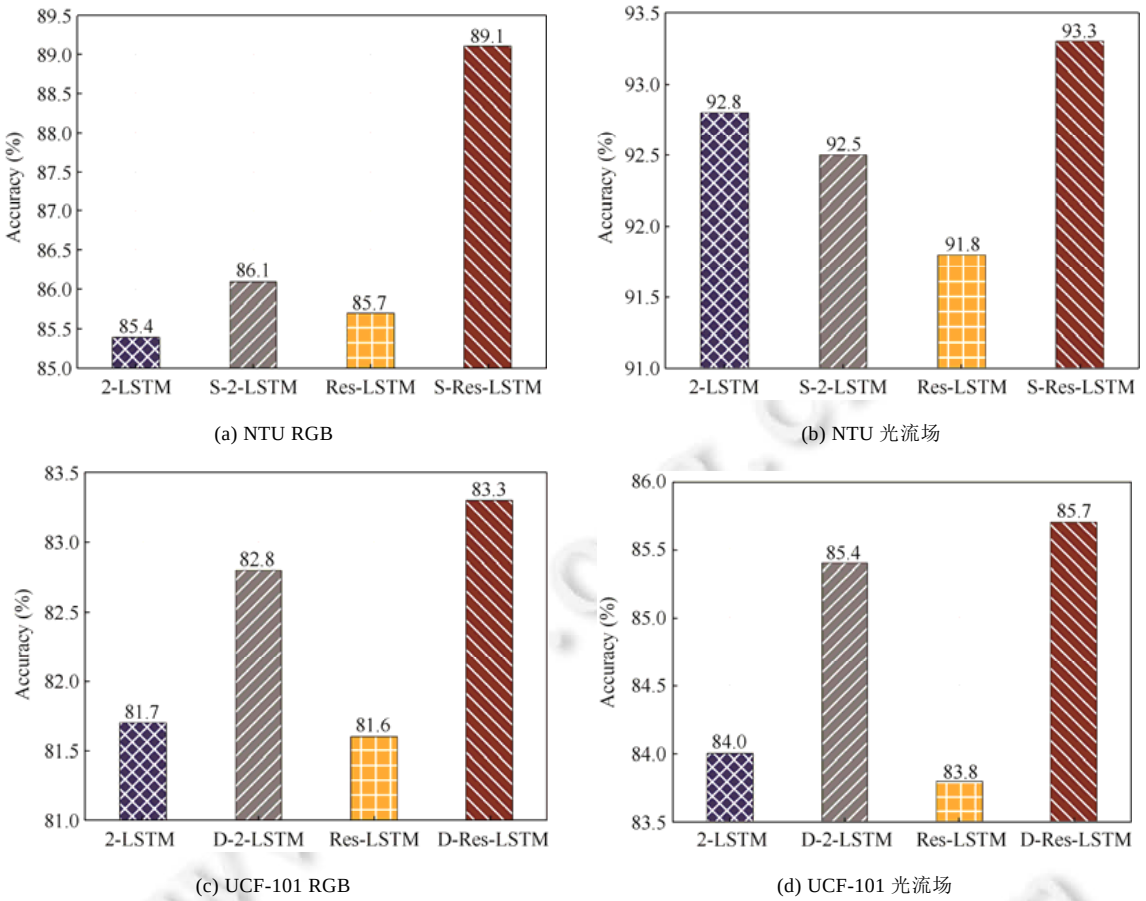


Fig.3 Performance comparisons with the structures of Res-LSTM and 2-LSTM on the NTU (CV) and UCF-101 (split1) datasets, respectively

图 3 Res-LSTM 和 2-LSTM 结构在 NTU RGB+D(CV)和 UCF-101(split1)数据集上的性能比较

3.4 与现有算法的比较

表 4 和表 5 展示了本文基于关联模态补偿的动作识别算法的最终结果以及与其他现有算法的比较.在本文算法中,对一个视频的最终分类结果由平均融合空域支路和时域支路的概率分布得到[5].

Table 4 Comparisons with state-of-the-arts on the NTU RGB+D dataset in accuracy (%)

表 4 不同算法在 NTU RGB+D 数据集上的性能(准确率:%)

不同算法	3D 骨架	RGB	光流场	CS	CV
Trust Gate[13]	✔			69.2	77.7
STA-LSTM[11]	✔			73.4	81.2
VA-LSTM[43]	✔			79.4	87.6
Ske-LSTM	✔			70.4	83.4
ST-GCN [44]	✔			81.5	88.3
Deep STGC$_k$[45]	✔			74.9	86.3
P-CNN[42]		✔	✔	53.8	61.7
TSN[6]		✔	✔	88.5	90.4
Chained MT[14]	✔	✔	✔	80.8	–
Res-LSTM		✔	✔	88.8	94.1
S-Res-LSTM		✔	✔	89.5	95.2
S-Res-LSTM+Ske-LSTM	✔	✔	✔	**90.0**	**96.3**

Table 5 Comparisons with state-of-the-arts on the UCF-101 dataset in accuracy (%)
表 5 不同算法在 UCF-101 数据集上的性能(准确率:%)

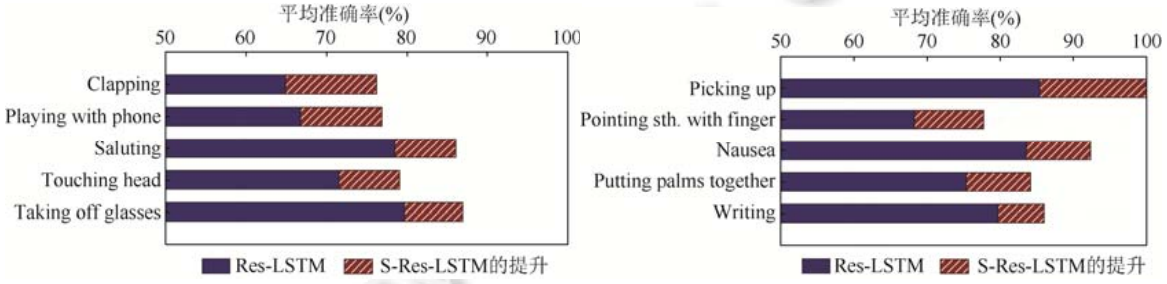
不同算法	准确率(%)
LRCN[26]	82.9
C3D[20]	85.2
Two-Stream ConvNet[3]	88.0
Two-Stream+Fusion[5]	92.5
TSN[6]	**94.0**
Chained MT[14]	91.3
NOASSOM[46]	92.1
T-C3D(Fast)[47]	91.8
T-C3D(Kinetics)[47]	92.5
Res-LSTM	91.7
D-Res-LSTM	92.3

NTU RGB+D 数据集被广泛用于评测基于 3D 骨架数据的动作识别算法.为了公平评价各种算法,表 4 中标注了不同算法所涉及到的数据模态.表 4 的 Ske-LSTM 展示了仅使用 1 层长短期记忆网络对 3D 骨架数据的识别结果,并对 RGB、光流场和 3D 骨架 3 个支路的概率进行平均融合,得到 S-Res-LSTM+Ske-LSTM 的结果.本文算法的关联模态适应能够使源模态数据吸取 3D 骨架数据的优点.在使用相同模态数据的前提下,本文算法在 NTU RGB+D 数据集上的准确率远高于其他现有算法.其中,本文通过执行 P-CNN[42]和 TSN[6]公开的源代码得到对应的动作识别准确率,文献[14]中只提供了 NTU RGB+D 数据集上 CV 场景下的结果.此外,表 5 中展示了 UCF-101 数据集上不同算法在 3 种划分方式下的平均实验性能.本文算法 D-Res-LSTM 取得了与现有算法相近的性能.

3.5 分析与讨论

分类分析:图 4 展示了相比于 Res-LSTM,在 NTU RGB+D 数据集(CV)中,平均准确率被 S-Res-LSTM 提升最多的 5 类动作.对于 RGB 数据来讲,提升最多的动作大多与运动细节相关,如鼓掌(clapping)、敬礼(saluting)等,说明在 3D 骨架的辅助下,本文算法可有效地对 RGB 特征补充运动细节信息.另一方面,对于光流场数据,一些动作的光流信息不太明显,降低了动作识别的准确度.而经过模态适应后,光流数据可在一定程度上传达形态信息,提升这类动作的识别准确率,如写字(writing).

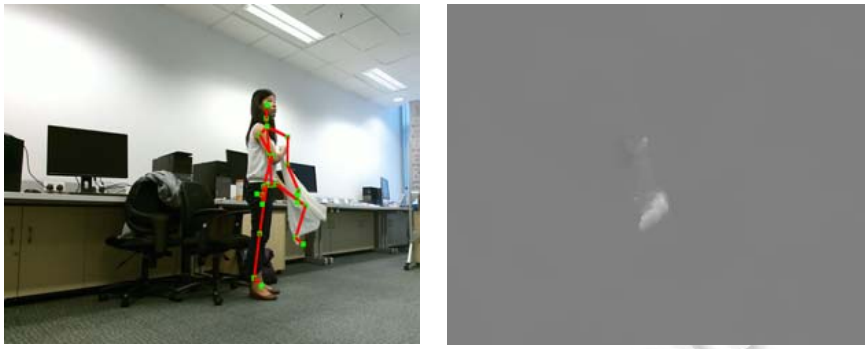
本文同时分析了 S-Res-LSTM 相比于 Res-LSTM 性能退化的情况,在图 5 中列举了本文方法性能不好的实例.在针对 RGB 图像的分类中,S-Res-LSTM 将图 5(a)所示识别为了“脱下外套”,主要原因为在“穿上外套”的动作中,3D 骨架由于遮挡等问题估计有误,导致 3D 骨架对“脱下外套”和“穿上外套”两类动作因噪声问题而区分性不强,从而影响了 S-Res-LSTM 对 RGB 的分类性能.另一方面,在针对光流的分类中,S-Res-LSTM 将图 5(b)所示“扇扇子”识别为了“挥手”,主要由于这两种动作在 3D 骨架在手部的运动上比较相近,而光流能够同时体现扇子和手部的运动,因此,S-Res-LSTM 在这类动作上存在退化.



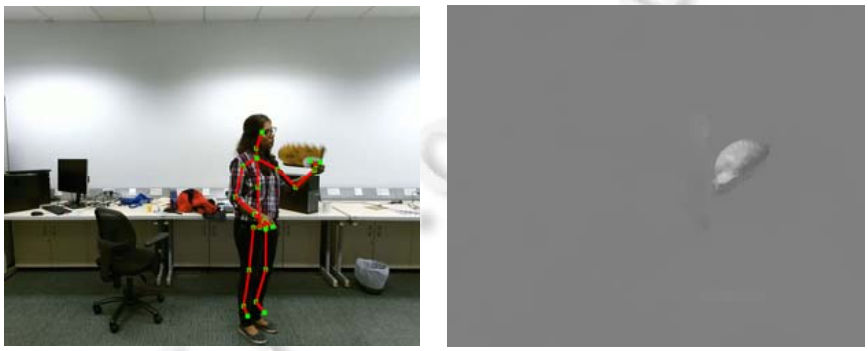
(a) NTU RGB (b) NTU 光流场

Fig.4 Top 5 activity categories in the NTU RGB+D (CV) dataset for which the recognition performance is improved the most by S-Res-LSTM, compared with Res-LSTM

图 4 相比于 Res-LSTM,NTU RGB+D(CV)数据集中被 S-Res-LSTM 提升最高的 5 类动作



(a) 对 RGB 的识别结果:Res-LSTM:脱下外套,S-Res-LSTM:穿上外套(正确类别:脱下外套)

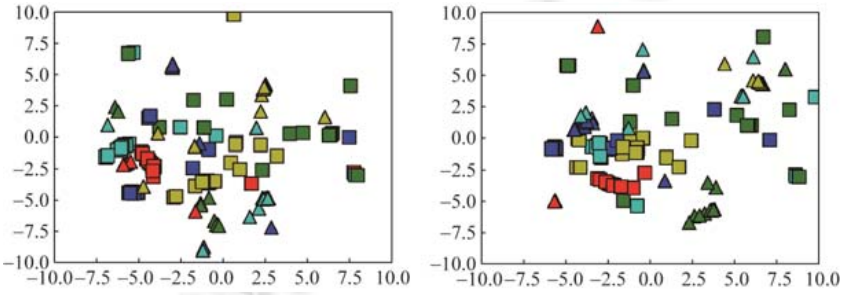


(b) 对光流的识别结果:Res-LSTM:扇扇子,S-Res-LSTM:挥手(正确类别:扇扇子)

Fig.5 Failure cases for S-Res-LSTM, compared with Res-LSTM

图 5 S-Res-LSTM 相比于 Res-LSTM 的退化情形

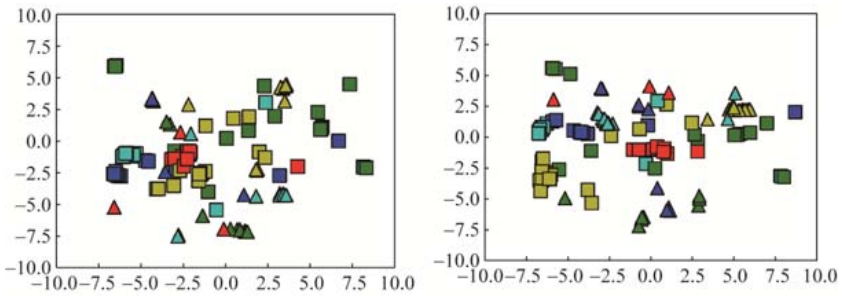
可视化分析:为了更好地理解损失函数(见式(11))的意义,本文使用 t-SNE 技术对不同模态层次适应后的特征进行可视化,即在式(11)中使用不同的 d.本文随机挑选了 5 类动作的源模态和辅助模态的特征,如图 6 所示,其中,不同的类别用颜色区分,源模态特征由方形表示,辅助模态特征由三角形表示,图中第 1 排的源模态为 RGB,第 2 排的源模态为光流.结果显示,在损失函数中使用模态层次或者类别层次的约束项时,源模态和辅助模态特征空间整体变得更加紧密.当使用样本层次的约束项时,同一类别源模态和辅助模态特征更具一致性.可视化分析表明,本文提出的模态适应能够有效地增强源模态特征的判别力.



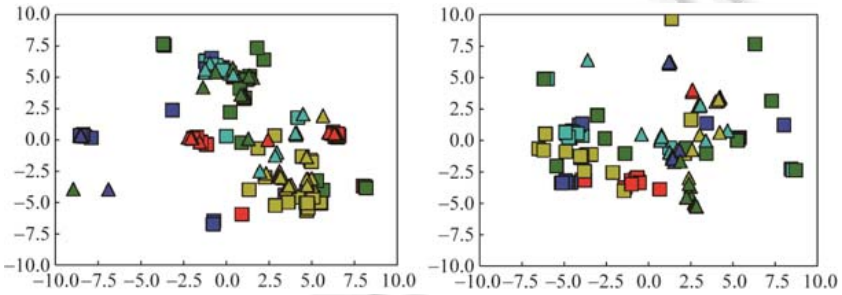
(a) d=0(无模态适应)

Fig.6 Visualization of adapted features with different d in the loss function

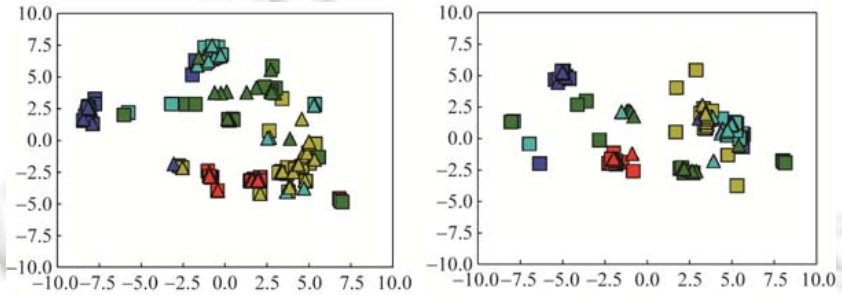
图 6 在损失函数中使用不同 d 后,对模态适应后特征的可视化



(b) $d=d_D$(模态层次)



(c) $d=d_C$(类别层次)



(d) $d=d_S$(样本层次)

Fig.6 Visualization of adapted features with different d in the loss function (Continued)

图 6 在损失函数中使用不同 d 后,对模态适应后特征的可视化(续)

4 结束语

本文提出了基于关联模态补偿的视频动作识别算法.该算法利用辅助模态数据和源模态数据的关联性,借助辅助模态的优势,对源模态数据做出特征补偿.其中,基于残差子网络的模态适应模块通过统一源模态和辅助模态特征空间的数据分布,在维持原有信息的同时,使源模态特征有能力表达辅助模态特征的优势.为了适用于对齐程度不同的训练数据,充分挖掘多模态数据的关联互补关系,本文提出了多层次模态适应方法.实验部分验证了基于关联模态补偿的动作识别算法中每一部分的有效性,并在通用公共数据集 NTU RGB+D 和 UCF-101 上实现了优越的视频动作识别性能.

References:

[1] Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding, 2011,115(2):224–241.

[2] Wang H, Kläser A, Schmid C, Liu C. Action recognition by dense trajectories. In: Proc. of the IEEE Computer Vision and Pattern Recognition. 2011. 3169–3176.

[3] Simonyan K, Zisserman A. Two-Stream convolutional networks for action recognition in videos. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 568–576.

[4] Feichtenhofer C, Pinz A, Wildes RP. Spatiotemporal multiplier networks for video action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7445–7454.

[5] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proc. of the IEEE Computer Vision and Pattern Recognition. 2016. 1933–1941.

[6] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L. Temporal segment networks: Towards good practices for deep action recognition. In: Proc. of the European Conf. on Computer Vision. 2016. 20–36.

[7] Wang Y, Long M, Wang J, Philip SY. Spatiotemporal pyramid network for video action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2097–2106.

[8] Zhang Z. Microsoft kinect sensor and its effect. IEEE Multimedia, 2012,19(2):4–10.

[9] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7291–7299.

[10] Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X. Co-Occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2016,2:8.

[11] Song S, Lan C, Xing J, Zeng W, Liu J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2017,1(2):7.

[12] Shahroudy A, Liu J, Ng TT, Wang G. NTU RGB+ D: A large scale dataset for 3D human activity analysis. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 1010–1019.

[13] Liu J, Shahroudy A, Xu D, Wang G. Spatio-Temporal LSTM with trust gates for 3D human action recognition. In: Proc. of the European Conf. on Computer Vision. 2016. 816–833.

[14] Zolfaghari M, Oliveira GL, Sedaghat N, Brox T. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2923–2932.

[15] Shi Z, Kim TK. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3461–3470.

[16] Mahasseni B, Todorovic S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3054–3062.

[17] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification. In: Proc. of the European Conf. on Computer Vision. 2010. 143–156.

[18] Szegedy C, Liu W, Jia Y, *et al*. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1–9.

[19] He K, Zhang X, Ren S, Sun, J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770–778.

[20] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4489–4497.

[21] Li Y, Li W, Mahadevan V, Vasconcelos N. VLAD3: Encoding dynamics of deep features for action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 1951–1960.

[22] Fernando B, Anderson P, Hutter M, Gould S. Discriminative hierarchical rank pooling for activity recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 1924–1932.

[23] Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S. Dynamic image networks for action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3034–3042.

[24] Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 4694–4702.

[25] Wu Z, Jiang YG, Wang X, Ye H, Xue X. Multi-Stream multi-class fusion of deep networks for video classification. In: Proc. of the ACM on Multimedia Conf. 2016. 791–800.

[26] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-Term recurrent convolutional networks for visual recognition and description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2625–2634.

[27] Lehrmann AM, Gehler PV, Nowozin S. Efficient nonlinear Markov models for human motion. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 1314–1321.

[28] Wu D, Shao L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 724–731.

[29] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 588–595.

[30] Vemulapalli R, Chellapa R. Rolling rotations for recognizing human actions from 3D skeletal data. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4471–4479.

[31] Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 1290–1297.

[32] Ke Q, Bennamoun M, An S, Sohel F, Boussaid F. A new representation of skeleton sequences for 3D action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4570–4579.

[33] Wang A, Cai J, Lu J, Cham TJ. MMSS: Multi-Modal sharable and specific feature learning for RGB-D object recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1125–1133.

[34] Wang J, Wang Z, Tao D, See S, Wang G. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: Proc. of the European Conf. on Computer Vision. 2016. 664–679.

[35] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010,22(10):1345–1359.

[36] Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D. Domain separation networks. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 343–351.

[37] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 448–456.

[38] Graves A. Supervised Sequence Labelling with Recurrent Neural Networks. Springer Science, Business Media, 2012. 385.

[39] Soomro K, ZamirA R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv Preprint, arXiv:1212.0402, 2012.

[40] Kingma DP, Ba J. ADAM: A method for stochastic optimization. arXiv Preprint, arXiv:1412.6980, 2014.

[41] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv Preprint, arXiv:1409.2329, 2014.

[42] Chéron G, Laptev I, Schmid C. P-CNN: Pose-Based CNN features for action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3218–3226.

[43] Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2117–2126.

[44] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. 7444–7452.

[45] Li C, Cui Z, Zheng W, Xu C, Yang J. Spatio-Temporal graph convolution for skeleton based action recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. 3482–3489.

[46] Du Y, Yuan C, Hu W, Yang H. Hierarchical nonlinear orthogonal adaptive-subspace self-organizing map based feature extraction for human action recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. 6805–6812.

[47] Liu K, Liu W, Gan, C, Tan M, Ma H. T-C3D: Temporal convolutional 3D network for real-time action recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. 7138–7145.

[48] Li Z, Tang J, He X. Robust structured nonnegative matrix factorization for image representation. IEEE Trans. on Neural Networks and Learning Systems, 2018,29(5):1947–1960.

[49] Li Z, Tang J. Weakly supervised deep matrix factorization for social image understanding. IEEE Trans. on Image Processing, 2017,26(1):276–288.

[50] Schroder M, Ritter H. Hand-Object interaction detection with fully convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2017. 18–25.

[51] Xu D, Ouyang W, Ricci E, Wang X, Sebe N. Learning cross-modal deep representations for robust pedestrian detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 5363–5371.

[52] Kiela D, Grave E, Joulin A, Mikolov T. Efficient large-scale multi-modal classification. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. 5198–5204.

宋思捷(1994－),女,河南濮阳人,学士,主要研究领域为动作识别,图像处理.

刘家瑛(1983－),女,博士,副教授,CCF 高级会员,主要研究领域为图像/视频压缩编码,增强重建与分析理解.

厉扬豪(1993－),男,硕士,主要研究领域为动作识别,深度学习.

郭宗明(1966－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为图像处理,视频处理,多媒体通信.