



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
07
VOL.26

ISSN1006-8961
CN11-3758/TB



中国多媒体大会(ChinaMM2020)专刊



第26卷第7期（总第303期）

2021年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院空天信息创新研究院

中国图象图形学学会

北京应用物理与计算数学研究所

主 编 吴一戎

编辑出版 《中国图象图形学报》编辑出版委员会

通信地址 北京市海淀区北四环西路19号

邮 编 100190

电子信箱 jig@aircas.ac.cn

电 话 010-58887035

网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司

（邮政信箱：北京399信箱 邮编：100048）

印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Aerospace Information Research Institute, CAS

China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong

Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics

Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China

Zip code 100190

E-mail jig@aircas.ac.cn

Telephone 010-58887035

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation

(P.O.Box 399, Beijing 100048, P.R.China)

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

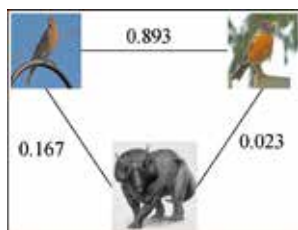
ISSN 1006-8961

CODEN ZTTXFZ

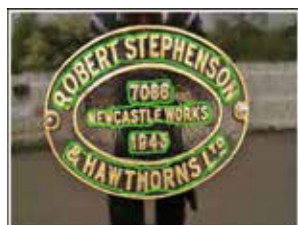
国外发行代号 M1406

国内邮发代号 82-831

国内定价 60.00元



类别语义相似性监督的小样本图像识别(第1594页)



像素聚合和特征增强的任意形状场景文本检测(第1614页)



时序特征融合的视频实例分割(第1692页)

综述

面向上行流媒体的压缩感知视频流技术前沿

刘浩, 黄荣, 袁浩东 1545

图像处理和编码

耦合保持投影哈希跨模态检索

闵康凌, 张国宾, 王磊, 李丹萍 1558

面向视觉搜索的空间局部敏感哈希方法

黄小燕, 孙彬, 杨展源, 朱映映, 田奇 1568

图像分析和识别

融合视觉关系检测的电力场景自动危险预警

高明, 左红群, 柏帆, 田清阳, 葛志峰, 董兴宁, 甘甜 1583

类别语义相似性监督的小样本图像识别

徐鹏帮, 桑基韬, 路冬媛 1594

局部双目视差回归的目标距离估计

张羽丰, 李昱希, 赵明璧, 喻晓源, 占云龙, 林巍峭 1604

像素聚合和特征增强的任意形状场景文本检测

师广琛, 巫义锐 1614

相位一致性指导的全参考全景图像质量评价

夏雨蒙, 王永芳, 王闯 1625

图像理解和计算机视觉

特征金字塔结构的时序行为识别网络

何嘉宇, 雷军, 李国辉 1637

多模态多层次事件网络的谣言检测

李莎, 张怀文, 钱胜胜, 方全, 徐常胜 1648

多模态零样本人体动作识别

吕露露, 黄毅, 高君宇, 杨小汕, 徐常胜 1658

足球视频球员感知跟踪算法

冯思佳, 宋子恺, 于俊清, 何云峰, 管涛 1668

融合时空图卷积的多人交互行为识别

成科扬, 吴金霞, 王文杉, 荣兰, 詹永照 1681

时序特征融合的视频实例分割

黄泽涛, 刘洋, 于成龙, 张加佳, 王轩, 漆舒汉 1692

医学图像处理

多尺度深度特征提取的肝脏肿瘤CT图像分类

毛静怡, 宋余庆, 刘哲 1704

3D多尺度深度卷积神经网络肺结节检测

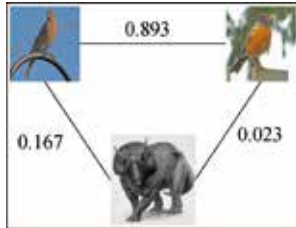
孙华聪, 彭延军, 郭燕飞, 张晓庆 1716

结合多通道注意力的糖尿病性视网膜病变分级

顾婷菲, 郝鹏翼, 白琮, 柳宁 1726

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Few shot image recognition based on class semantic similarity supervision(P1594)



Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement (P1614)



Video instance segmentation based on temporal feature fusion (P1692)

Review

Survey on compressive sensing video stream for uplink streaming media

Liu Hao, Huang Rong, Yuan Haodong 1545

Image Processing and Coding

Structure-preserving hashing with coupled projections for cross-modal retrieval

Min Kangling, Zhang Guobin, Wang Lei, Li Danping 1558

Locality-sensitive hashing approach based on semantic space for visual retrieval

Huang Xiaoyan, Sun Bin, Yang Zhanyuan, Zhu Yingying, Tian Qi 1568

Image Analysis and Recognition

Visual relationship detection-based emergency early-warning description generation in electric power industry

Gao Ming, Zuo Hongqun, Bai Fan, Tian Qingyang, Ge Zhifeng, Dong Xingning, Gan Tian 1583

Few shot image recognition based on class semantic similarity supervision

Xu Pengbang, Sang Jitao, Lu Dongyuan 1594

Object distance estimation based on stereo regional disparity regression

Zhang Yufeng, Li Yuxi, Zhao Mingbi, Yu Xiaoyuan, Zhan Yunlong, Lin Weiyao 1604

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement

Shi Guangchen, Wu Yirui 1614

Phase consistency guided full-reference panoramic image quality assessment algorithm

Xia Yumeng, Wang Yongfang, Wang Chuang 1625

Image Understanding and Computer Vision

Temporal action detection based on feature pyramid hierarchies

He Jiayu, Lei Jun, Li Guohui 1637

Multi-modal multi-level event network for rumor detection

Li Sha, Zhang Huaiwen, Qian Shengsheng, Fang Quan, Xu Changsheng 1648

Multimodal-based zero-shot human action recognition

Lyu Lulu, Huang Yi, Gao Junyu, Yang Xiaoshan, Xu Changsheng 1658

Players-aware tracking algorithm in soccer video

Feng Sijia, Song Zikai, Yu Junqing, He Yunfeng, Guan Tao 1668

Multi-person interaction action recognition based on spatio-temporal graph convolution

Cheng Keyang, Wu Jinxia, Wang Wenshan, Rong Lan, Zhan Yongzhao 1681

Video instance segmentation based on temporal feature fusion

Huang Zetao, Liu Yang, Yu Chenglong, Zhang Jiajia, Wang Xuan, Qi Shuhan 1692

Medical Image Processing

CT image classification of liver tumors based on multi-scale and deep feature extraction

Mao Jingyi, Song Yuqing, Liu Zhe 1704

3D multi-scale deep convolutional neural networks in pulmonary nodule detection

Sun Huacong, Peng Yanjun, Guo Yanfei, Zhang Xiaoqing 1716

Diabetic retinopathy grading based on multi-channel attention

Gu Tingfei, Hao Pengyi, Bai Cong, Liu Ning 1726

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2021)07-1681-11

论文引用格式: Cheng K Y, Wu J X, Wang W S, Rong L and Zhan Y Z. 2021. Multi-person interaction action recognition based on spatio-temporal graph convolution. Journal of Image and Graphics, 26(07): 1681-1691 (成科扬, 吴金霞, 王文杉, 荣兰, 詹永照. 2021. 融合时空图卷积的多人交互行为识别. 中国图象图形学报, 26(07): 1681-1691) [DOI:10.11834/jig.200510]

融合时空图卷积的多人交互行为识别

成科扬^{1,2,3}, 吴金霞^{1*}, 王文杉⁴, 荣兰¹, 詹永照¹

1. 江苏大学计算机科学与通信工程学院, 镇江 212013; 2. 江苏省大数据泛在感知与智能农业应用工程研究中心, 镇江 212013;
3. 江苏大学网络空间安全研究院, 镇江 212013; 4. 中国电子科学研究院社会安全风险感知与防控大数据
应用国家工程实验室, 北京 100041

摘要: 目的 多人交互行为的识别在现实生活中有着广泛应用。现有的关于人类活动分析的研究主要集中在对单人简单行为的视频片段进行分类, 而对于理解具有多人之间关系的复杂人类活动的问题还没有得到充分的解决。**方法** 针对多人交互动作中两人肢体行为的特点, 本文提出基于骨架的时空建模方法, 将时空建模特征输入到广义图卷积中进行特征学习, 通过谱图卷积的高阶快速切比雪夫多项式进行逼近。同时对骨架之间的交互信息进行设计, 通过捕获这种额外的交互信息增加动作识别的准确性。为增强时域信息的提取, 创新性地将切片循环神经网络(recurrent neural network, RNN)应用于视频动作识别, 以捕获整个动作序列依赖性信息。**结果** 本文在UT-Interaction数据集和SBU数据集上对本文算法进行评估, 在UT-Interaction数据集中, 与H-LSTCM(hierarchical long short-term concurrent memory)等算法进行了比较, 相较于次好算法提高了0.7%, 在SBU数据集中, 相较于GCNConv(semi-supervised classification with graph convolutional networks)、RotClips+MTCNN(rotating cliips+multi-task convolutional neural network)、SGC(simplifying graph convolutional)等算法分别提升了5.2%、1.03%、1.2%。同时也在SBU数据集中进行了融合实验, 分别验证了不同连接与切片RNN的有效性。**结论** 本文提出的融合时空图卷积的交互识别方法, 对于交互类动作的识别具有较高的准确率, 普遍适用于对象之间产生互动的行为识别。**关键词:** 动作识别; 交互信息; 时空建模; 图卷积; 切片循环神经网络(RNN)

Multi-person interaction action recognition based on spatio-temporal graph convolution

Cheng Keyang^{1,2,3}, Wu Jinxia^{1*}, Wang Wenshan⁴, Rong Lan¹, Zhan Yongzhao¹

1. School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang 212013, China;
2. Jiangsu Province Big Data Ubiquitous Perception and Intelligent Agricultural Application Engineering Research Center, Zhenjiang 212013, China; 3. Cyber Space Security Academy of Jiangsu University, Zhenjiang 212013, China;
4. National Engineering Laboratory for Public Security Risk Perception and Control by Big Data, China Academy of Electronic Sciences, Beijing 100041, China

Abstract: **Objective** The recognition of multi-person interaction behavior has wide applications in real life. At present, human activity analysis research mainly focuses on classifying video clips of behaviors of individual persons, but the prob-

收稿日期: 2020-08-23; 修回日期: 2020-12-30; 预印本日期: 2021-01-06

* 通信作者: 吴金霞 2061383373@qq.com

基金项目: 国家自然科学基金项目(61972183); 社会安全风险感知与防控大数据应用国家工程实验室主任基金项目

Supported by: National Natural Science Foundation of China(61972183); National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC)

lem of understanding complex human activities with relationships between multiple people has not been resolved. When performing multi-person behavior recognition, the body information is more abundant and the description of the two-person action features are more complex. The problems such as complex recognition methods and low recognition accuracy occur easily. When the recognition object changes from a single person to multiple people, we not only need to pay attention to the action information of each person but also need to notice the interaction information between different subjects. At present, the interaction information of multiple people cannot be extracted well. To solve this problem effectively, we propose a multi-person interaction behavior-recognition algorithm based on skeleton graph convolution. **Method** The advantage of this method is that it can fully utilize the spatial and temporal dependence information between human joints. We design the interaction information between skeletons to discover the potential relationships between different individuals and different key points. By capturing the additional interaction information, we can improve the accuracy of action recognition. Considering the characteristics of multi-person interaction behavior, this study proposes a spatio-temporal graph convolution model based on skeleton. In terms of space, we have various designs for single-person and multi-person connections. We design the single-person connection within each frame. Apart from the physical connections between the points of the body, some potential correlations are also added between joints that represent non-physical connections such as the left and right hands of a single person. We design the interaction connection between two people within each frame. We use Euclidean distance to measure the correlation between interaction nodes and determine which points between the two persons have a certain connection. Through this method, the connection of the key points between the two persons in the frame not only can add new and necessary interaction connections, which can be used as a bridge to describe the interaction information of the two persons' actions, but can also prevent noise connections and cause the underlying graph to have a certain sparseness. In the time dimension, we segment the action sequence. Every three frames of action are used as a processing unit. We design the joints between three adjacent frames, and use more adjacent joints to expand the receptive field to help us learn the change information in the time domain. Through the modeling design in the time and space dimensions, we have obtained a complex action skeleton diagram. We use the generalized graph convolution model to extract and summarize the two people action features, and approximate high-order fast Chebyshev polynomials of spectral graph convolution to obtain high-level feature maps. At the same time, to enhance the extraction of time domain information, we propose the application of sliced recurrent neural network (RNN) to video action recognition to enhance the characterization of two people actions. By dividing the input sequence into multiple equiling subsequences and using a separate RNN network for feature extraction on each subsequence, we can calculate each subsequence at the same time, thereby overcoming the limitations of sliced RNN that cannot be parallelized. Through the information transfer between layers, the local information on the subsequence can be integrated in the high-level network, which can integrate and summarize the information from local to global, and the network can capture the entire action-sequence dependent information. For the loss of information at the slice, we have solved this problem by taking the three frame actions as a processing unit. **Result** This study validates the proposed algorithm on two datasets (UT-Interaction and SBU) and compares them with other advanced interaction-recognition methods. The UT-Interaction dataset contains six classes of actions and the SBU interaction dataset has eight classes of actions. We use 10-fold and 5-fold cross-validation for evaluation. In the UT-Interaction dataset, compared with H-LSTCM (Chierarchical long-short-term concurrent memory) and other methods, the performance improves by 0.7% based on the second-best algorithm. In the SBU dataset, compared with GCNConv, RotClips + MTCNN, SGCCConv, and other methods, the algorithm has been improved by 5.2%, 1.03%, and 1.2% respectively. At the same time, fusion experiments are conducted in the SBU dataset to verify the effectiveness of various connections and sliced RNN. This method can effectively extract additional information on interactions, and has a good effect on the recognition of interaction actions. **Conclusion** In this paper, the interactive recognition method of fusion spatio-temporal graph convolution has high accuracy for the recognition of interactive actions, and it is generally applicable to the recognition of behaviors that generate interaction between objects.

Key words: action recognition; interaction information; spatial-temporal modeling; graph convolution; sliced recurrent neural network (RNN)

0 引言

基于视频的交互行为识别具有较高的实用价值和广阔的应用前景(王世刚等,2015)。人类动作识别的目的是分析和理解视频中人与人之间的动作和相互作用。它可以应用于智能监控、人机交互、视频序列理解和医疗卫生等多个领域,在日常生活中发挥着越来越重要的作用。

在行为识别中,人体骨架数据的应用相对于RGB数据和深度数据来说有着明显的优势,它可以在不受背景、光照和外貌等各种因素的影响,此外,骨架特征紧凑、结构性强、语义丰富,对人体动作和运动的描述能力强,越来越多的行为识别基于骨架进行研究。目前基于深度学习的骨架交互识别方法主要有以下3种:长短期记忆网络(long short-term memory, LSTM)、卷积神经网络(convolutional neural network, CNN)以及图卷积网络(graph convolution network, GCN)。

目前,比较成熟的研究都是针对单人骨架动作进行识别,缺少对交互动作的讨论。然而在日常生活中,常见的行为基本都是一些交互行为,如握手、拥抱、打斗等。相较于单人动作,交互类动作复杂程度更高,在完成交互动作的过程中,肢体动作种类更多,肢体间的变化也更加多元化。因此,如何有效提取交互动作的特征,并对交互行为进行建模和分析是极具挑战性的问题。

以往的工作是将骨架数据重新组织成一种网格状结构,通过RNN(recurrent neural network)和CNN来实现。虽然在动作识别方面有很大的改进,但由于人体骨架是图形结构,而不是传统的固定网格,因此并没有完全受益于深度学习的优越表示能力。人体骨架是在非欧几里德空间中自然构建的图形,CNN方法虽然具有强大的特征提取能力,但需要固定大小的卷积核进行遍历处理,因此不能有效提取图数据关键特征,而且其计算复杂度较大,在处理多人任务时不能满足其精度要求,这使得传统的卷积神经网络不适用。虽然传统RNN也可以处理骨架,但它直接对骨架数据进行转换,识别的精度不高,因此,本文通过GCN来处理转换后的骨架数据,用以捕获运动空间特征,并结合变种RNN结构来捕获时域依赖信息。GCN可以直接对原始骨架数据进行

建模,将图神经网络扩展到时空图模型中,自动地从骨架中学习空间和时间的信息。Yan等人(2018)将图卷积网络引入基于骨架的动作识别,并取得了许多令人鼓舞的结果。然而,大多数GCN方法都是基于具有固定拓扑约束的预定义图,忽略了隐含的关节关联。同时,GCN不能够完整地捕获整个动作序列的时间信息,且无法获得动作序列依赖性信息。

为了解决这个问题,本文设计了各种自适应的连接。强调个体、交互对象、时间帧之间的关系。同时增强时序依赖信息的提取。本文主要有以下3个贡献:

1)在交互动作识别时,通过对参与者身体各部分之间的相互作用关系建模,提取来自交互本身的附加信息,将这些信息用于识别人类交互的全局描述符中,以提高交互动作识别准确性。

2)提出了基于骨架的时空建模方法,不仅对单帧内的单个对象和多个对象分别进行连接设计,还结合单帧和多帧的不同连接,而且还与前一帧和后一帧中的相关关节连接,实现交互骨架图的有效表示。

3)创新性地将切片RNN应用于视频动作识别领域,以增强视频序列依赖性信息的提取。同时,时空建模方法与切片RNN进行结合有效地弥补切片RNN的弊端。

1 相关工作

1.1 基于图像的交互识别

早期的识别工作大多是基于手工构建的特征,如使用方向梯度直方图(histogram of oriented gradient, HOG)和光流方向信息直方图(histograms of oriented optical flow, HOF)(Van Gemeren等,2014)等提取静态信息的外观特征,或者使用光流提取动态信息的运动特征。较新的方法依赖于深度学习,Ke等人(2016)将深度学习网络用于交互行为识别,通过CNN提取出光流特征信息,然后送入到分类器中实现动作识别。

虽然基于RGB视频或光流的动作识别方法具有较高的性能,但仍存在一些问题。例如易受背景、光照和外观变化的影响,提取光流信息需要较高的计算代价。一些工作已经对骨骼数据进行提取,以避免从视频中直接学习交互模式。在一些单人动作

识别研究中,学者多通过人体骨架来进行动作识别。人体骨架可以很好地表现人体的运动,有利于分析人体的动作。一方面,骨架数据在背景噪声中具有固有的鲁棒性,为人体运动提供了抽象、高层次的特征。另一方面,与 RGB 数据相比,骨架数据量非常小,这使得本文可以设计一个更好的模型。因此,本文将基于骨架的动作识别从单人扩展到多人。

1.2 基于骨骼的交互识别

随着深度学习的发展,基于骨骼的方法逐渐出现。Liu 等人(2016)提出一个关节序列的时空 LSTM 网络,将 LSTM 的学习扩展到时空域,每个关节从相邻关节以及前一帧接受信息编码时空特征,采用树状结构表示关节之间的相邻特性和运动关系,最后将骨架数据的结果送入 LSTM 网络进行建模与识别。Du 等人(2015)根据人体的物理结构将人体骨骼分为 5 部分,分别将其分成 5 个双向递归连接的子网。Song 等人(2017)提出了一个端到端的时空注意力模型,用于从骨骼数据识别人类动作。在 LSTM 和 RNN 的基础上构建模型,设计了一个具有联合选择门的时空注意力模块,自适应地将不同的注意力分配到每个帧内输入框架的不同关节。也出现了一些基于 CNN 的方法,如 Liu 等人(2017)提出了一种增强的骨骼可视化方法,将骨骼序列表示为一系列增强的视觉和运动彩色图像,该方法以紧凑而独特的方式隐式地描述了时空骨骼关节。研究也将卷积神经网络与递归神经网络相结合,以对交互作用进行更复杂的时间推理(Donahue 等,2015)。Wang 和 Deng(2020)针对 RNN 和 CNN 在基于骨架的动作识别中表现出的良好性能,提出了一种将 CNN 分类与 RNN 相结合的深度网络结构,实现了

人体交互识别的注意机制。虽然基于 RNN 的方法具有很强的序列数据建模能力,并且基于 CNN 的方法具有很好的并行性,训练过程也比较简单,但是无论是 CNN 还是 RNN 都不能完全代表骨架的结构。Ke 等人(2018)提出了一种新的 3 维动作识别骨架序列表示方法,很好地捕捉到复杂的空间结构和骨架序列的长期时间动态。与 RNN 和 CNN 的方法相比,基于 GCN 的方法能够更好地表达节点之间的依赖关系。Shi 等人(2019)将非局部神经网络引入到图卷积运算中,对多层次的语义信息进行建模,具有更大的灵活性和通用性。Tang 等人(2018)提出了一种深度渐进强化学习(deep progressive reinforcement learning, DPRL)方法来提取关键帧,并利用基于图的卷积神经网络来捕捉关节之间的依赖关系进行动作识别。Gao 等人(2019)提出了一种基于图回归 GCN(graph regression-graph convolution network, GR-GCN)方法用于基于骨架的动作识别,以捕获数据中的时空变化。但是这些方法在识别交互的动作中没有显式的图构造,本文进一步利用骨架间的关系,提取人体之间的交互特征,并将图卷积与 RNN 相结合,更好地提取节点之间以及帧之间的依赖性信息。

2 交互信息的时空图构建

将关键点的连接分为帧内单人连接和交互连接以及帧间连接。如图 1 所示,通过不同的方法设计这些连接,随后运用谱图卷积得到其变化特征,并结合切片 RNN 获取时序依赖信息,进行动作识别。

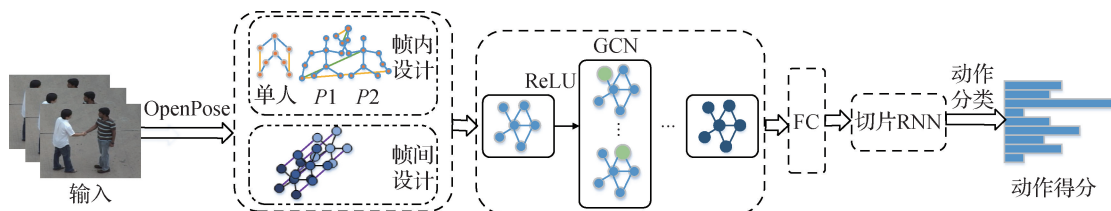


图 1 总体流程图

Fig. 1 Overall flow chart

2.1 帧内交互建模

帧内设计分为单人设计和交互设计。对于每一帧中的单个人,通过连通图来建模人体,人体连通图

仅通过个人的自然连接来表示,这样无法很好地提取出人体的全局信息,因此,通过关节之间的不同相关性进行连接,将其分为内部连接与外部连接。

内部连接包括关节之间的物理连接,外部连接表示非物理连接的关节之间的潜在关联。在拳击比赛中,虽然人体手部与头部不存在骨骼连接,但由于选手一般将双手架于面部之前,以作防守,因此双手与头部之间存在潜在关系,在它们之间建立外部连接。在加权邻接矩阵中设置不同的参数来区分这两种关系。如图2(a)所示,对内连接的边和外连接的边赋予不同的权重,帧内的边权重设置为

$$w_{i,j} = \begin{cases} \alpha & (i,j) \in \epsilon_1 \\ \beta & (i,j) \in \epsilon_2 \\ 0 & i = j \end{cases} \quad (1)$$

设置 $w_{i,j} = 0$ 表示关节之间没有连接。 α 和 β 表示内外连接时所设定的参数。此外,将关节之间连接分别用 ϵ_1 和 ϵ_2 来表示。 ϵ_1 表示关节之间的内在连接,如图2(a)的黑色实线所示,作为一个重要的性质,连接节点之间的距离在运动过程中保持不变。 ϵ_2 表示关节之间的外在连接,如图2(a)的虚线所示,外部依赖关系是指两个关节之间的断开关系,这也是动作过程中的一个重要因素。

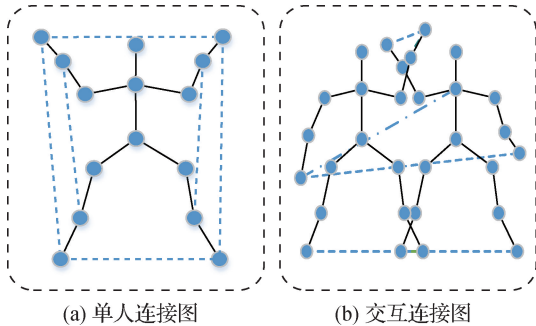


图2 帧内连通图

Fig. 2 In-frame connected graph((a) single connection diagram;(b) interactive connection diagram)

与以前的工作不同,在骨骼交互动作识别中,两个人关节之间的连接都是断开的,学习描述每个对象如何相互关联十分必要,需要将两个人及其交互信息进行合并。通过对两人之间骨骼的构造来提取其交互信息。对动作参与者之间进行交互设计,将两个独立的骨骼图通过关节点连接,整合成带有交互信息的动作骨骼图,通过图卷积网络便能够提取动作的交互信息。

交互设计包括两个部分,将易于出现类似关节变化的点的连接称为对应连接,用 ϵ_3 来表示对应连接,例如拥抱动作,两个参与者的动作基本一致,将

相对应的关节点之间建立连接,如图2(b)的虚线所示,当参与者发生的动作基本一致的时候,这些对应连接边的建立起到重要作用。另外,将其他关节点之间发生的连接称为潜在连接。如图2(b)的点划线所示,用 ϵ_4 来表示潜在连接。将 θ 指定给 ϵ_3 中的边的权重,将 δ 指定给 ϵ_4 中的边的权重,即

$$w_{i,j} = \begin{cases} \theta & (i,j) \in \epsilon_3 \\ \delta & (i,j) \in \epsilon_4 \\ 0 & \text{其他} \end{cases} \quad (2)$$

式中, i 和 j 表示不同人的关节点,将单帧内的邻接矩阵表示为

$$A_* = \begin{bmatrix} P_1 & M_1 \\ M_2 & P_2 \end{bmatrix} \quad (3)$$

式中, P_1 、 P_2 描述单人连接部分, M_1 、 M_2 描述交互连接部分。

为确定上述帧内交互建模哪些关节点进行相应的连接,通过欧氏距离来衡量交互类节点之间的相关性。计算所有点之间的值为

$$d(x_i, x_j) = \|x_i - x_j\|^2 \quad (4)$$

式中, x_i 、 x_j 分别为关键点 i 和关键点 j 的特征表示。仅计算外在连接和潜在连接的边的欧氏距离,然后,对所得出的距离进行规范化,将其结果映射到 $[0,1]$ 之间,采用最大最小值的归一化方法,即

$$n = \frac{d_i - d_{\min}}{d_{\max} - d_{\min}} \quad (5)$$

式中, $d_{\max}$ 表示关节距离最大值, $d_{\min}$ 表示关节距离最小值。

本文实验设定 $n < 0.3$ 时,生成一条新边连接,这样既可以添加一些新的必要的交互连接,也可以使底层图具有一定的稀疏性。

2.2 帧间建模

在时域中每个关节都是断开的,允许将帧 x_t 中的每个关节连接到其在前一帧 x_{t-1} 和后一帧 x_{t+1} 中对应的邻域,如图3所示。

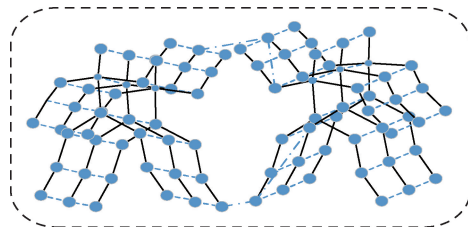


图3 帧间连通图

Fig. 3 A connected graph between frames

为了利用更多的相邻关节来扩大感受野,以此帮助学习时域的变化信息。与帧内连接相似,帧间的连接也有两种连接类型:对应关节连接,表示为 ε_5 ;以及相邻帧中每个关节与其对应关节的邻域之间的连通连接,表示为 ε_6 。这两种边的权重表示为

$$w_{i,j} = \begin{cases} \gamma & (i,j) \in \varepsilon_5 \\ \lambda & (i,j) \in \varepsilon_6 \\ 0 & \text{其他} \end{cases} \quad (6)$$

式中, i 和 j 表示不同帧之间的关节点。将最终构造的多帧邻接矩阵表示为

$$\mathbf{A}_{\text{total}} = \begin{bmatrix} \mathbf{A}_{*(t-1)} & \mathbf{A}_{t-1,t} & \mathbf{0} \\ \mathbf{A}_{t,t-1} & \mathbf{A}_{*(t)} & \mathbf{A}_{t,t+1} \\ \mathbf{0} & \mathbf{A}_{t+1,t} & \mathbf{A}_{*(t+1)} \end{bmatrix} \quad (7)$$

式中, $\mathbf{A}_{*(i)}$ 表示帧 i 的帧内建模图的邻接矩阵, $\mathbf{A}_{i,j}$ 表示帧 i 与帧 j 之间的邻接矩阵。 $\mathbf{0}$ 表示零矩阵。因此计算的图拉普拉斯: $\mathbf{L} = \mathbf{D} - \mathbf{A}_{\text{total}}$ 。

2.3 基于连通图的谱图卷积算法

以关节为节点,节点之间的连接为边来构造骨架图。在一个框架中,关节通过内部与外部进行连接充当空间边缘,帧间连接充当时间边缘,每个节点的属性是关节的坐标向量。将谱图卷积运算应用于时空骨架图以获得高级特征图。

考虑一个无向图 $\mathbf{G} = \{\mathbf{V}, \mathbf{E}, \mathbf{A}\}$ 由顶点集 $\mathbf{V}$ 、连接顶点的边集 $\mathbf{E}$ 和加权邻接矩阵 $\mathbf{A}$ 组成。 $\mathbf{A}$ 是实对称矩阵, $a_{i,j}$ 是分配给连接顶点 i 和 j 的边 (i,j) 的权重。假设权重非负,即 $a_{i,j} \geq 0$ 。由邻接矩阵定义的拉普拉斯矩阵可以用来揭示图的许多有用性质。在拉普拉斯矩阵的不同变体中,使用的组合图拉普拉斯定义为

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \quad (8)$$

式中, $\mathbf{D}$ 是 $d_{ii} = \sum_{j=1}^n a_{i,j}$ 的度矩阵,对称归一化拉普拉斯定义为 $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}}$ 。

基于骨架的动作识别的基础是捕捉关节的变化,从而学习运动特征进行分类。通过图拉普拉斯来模拟骨骼的变化。拉普拉斯矩阵 $\mathbf{L}$ 本质上是一个高通算子,能够捕捉底层信号的变化。对于任何信号 $\mathbf{x} \in \mathbf{R}^N$,满足

$$\mathbf{Lx}_i = \sum_{j \in N_i} a_{i,j} (\mathbf{x}_i - \mathbf{x}_j) \quad (9)$$

式中, $\mathbf{Lx}_i$ 表示 $\mathbf{Lx}$ 的第 i 个分量, $\mathbf{Lx}$ 表示拉普拉斯

矩阵作用于 $\mathbf{x}$ 上。 N_i 是连接到 i 的顶点集。

采用切比雪夫多项式近似地表达谱图卷积的结果,即

$$\mathbf{g}_\theta * \mathbf{x} \approx \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\mathbf{L}}) \mathbf{x} \quad (10)$$

式中, $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}}$ 是定义的对称归一化图拉普拉斯。 θ_k 表示第 k 个切比雪夫系数, $\mathbf{g}_\theta$ 表示卷积核。 $T_k(\tilde{\mathbf{L}})$ 是 k 阶的切比雪夫多项式。由 $T_k(\tilde{\mathbf{L}}) = 2\tilde{\mathbf{L}}T_{k-1}(\tilde{\mathbf{L}}) - T_{k-2}(\tilde{\mathbf{L}})$ 反复计算得到, $T_0(\tilde{\mathbf{L}}) = 1$, $T_1(\tilde{\mathbf{L}}) = \tilde{\mathbf{L}}$ 。因此,1 阶切比雪夫多项式计算 $(\tilde{\mathbf{L}})\mathbf{x}$,正好对应于式(9),从而捕获骨架数据中的变化。当 $k > 1$ 时, $\tilde{\mathbf{L}}^k$ 表达了当前节点与周围“ k 跳”邻居节点的卷积信息。

因此,图卷积定义为

$$y = \text{ReLU}\left(\sum_{k=0}^{K-1} T_k(\tilde{\mathbf{L}}) \mathbf{x} \mathbf{W}_k + b\right) \quad (11)$$

式中, $\mathbf{W}_k \in \mathbf{R}^{F_1 \times F_2}$ 是网络中学习的权重参数 θ'_k 的矩阵, F_1, F_2 分别是两个连接层中生成特征的维数。 b 为偏置,ReLU 为激活函数。在图数据卷积层之后,将 GCN 提取的空间特征序列输入切片 RNN 网络层,进行时序特征提取。为了使输入序列长度适应切片 RNN 的输入要求,通过一个全连接层调整数据维度。最后通过 softmax 激活函数生成输出分类。

2.4 基于切片 RNN 的时序建模

2.2 节已经对帧间进行建模,以扩大感受野、学习时域变化信息,但是这种帧间建模并不能够完整地捕获整个动作序列的时间信息,且无法获得动作序列依赖性信息。因此,在时间序列的处理上采用 RNN 进行处理,以解决动作序列数据的依赖性问题。但是传统 RNN 网络的当前节点信息只与前一节点相关,导致其只能够建模短时间的动态信息,无法存储长时序;同时标准的 RNN 网络结构无法实现像 CNN 网络模型的并行计算,因此采用切片 RNN 网络模型以解决上述问题。切片 RNN 结构模型如图 4 所示。

如图 4 所示,将输入序列划分成多个序列片段,在每一个片段上采用独立的 RNN 网络进行计算,本文 RNN 隐藏单元采用门控循环单元(gate recurrent unit,GRU),这样不仅实现了计算的“并行性”,而且在每一个相对较短的序列片段上进行 RNN 特征提取,通过层之间的信息传递能够更大程度地保留长期依赖性的信息。 $\mathbf{H}$ 表示网络的隐层状态, $\mathbf{Y}$ 表示

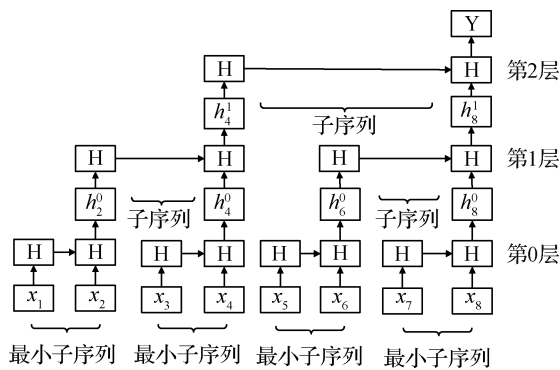


图4 切片RNN结构

Fig. 4 Section RNN structure

顶层输出。对于切片处造成的长期依赖性损失(Li等,2019),通过帧间建模,使得输入数据本身就能够弥补这一损失。

切片RNN输入序列表示为 $\mathbf{X} = [x_1, x_2, \dots, x_T]$, T 为序列长度,将 $\mathbf{X}$ 划分为 n 个等长,那么每个子序列 N 的长度 t 为

$$t = \frac{T}{n} \quad (12)$$

从而输入表示为 $\mathbf{X} = [N_1, N_2, \dots, N_n]$, $N_p = [x_{(p-1)t+1}, x_{(p-1)t+2}, \dots, x_{pt}]$ 。同样,将每个子序列 N 再次切片成 n 个等长的子序列,然后重复这个切片操作 k 次,直到在底层(称之为第0层,如图3所示)有一个适当的最小子序列长度,通过切片 k 次得到 $k+1$ 层。第0层最小子序列长度和最小子序列个数分别为

$$l_0 = \frac{T}{n^k} \quad (13)$$

$$s_0 = n^k \quad (14)$$

由于在第 p 层($p > 0$)的每个父序列被切片为 n 块,因此第 p 层的子序列个数和子序列长度分别为

$$s_p = n^{k-p} \quad (15)$$

$$l_p = n \quad (16)$$

在第0层,递归单元通过连接结构作用于每个最小子序列上。然后,得到第0层上每个最小子序列的最后一个隐藏状态,并将其作为第1层父序列的输入。然后利用第 $p-1$ 层上每个子序列的最后一个隐藏状态作为其父序列在第 p 层的输入,并计算出第 p 层上子序列的最后一个隐藏状态

$$h_t^1 = GRU^0(mss_{(t-l_0+1) \sim t}^0) \quad (17)$$

$$h_t^{p+1} = GRU^p(h_{t-l_p}^p \sim h_t^p) \quad (18)$$

式中, h_t^p 表示第 p 层的第 t 个子序列的隐层表示。

mss^0 表示第0层上的最小子序列, $mss_{(t-l_0+1) \sim t}^0$ 为在第0层上子序列内隐藏状态的计算,不同的GRU可以用于不同的层。式(18)表示在第0层计算出隐藏状态后,利用计算结果再次计算下一个隐藏状态,重复计算。这个操作在每个层上的每个子序列之间重复,直到得到顶层(第 k 层)的最终隐藏状态

$$\mathbf{F} = GRU^k(h_{t-l_k}^k \sim h_t^k) \quad (19)$$

与标准RNN类似,softmax层在最终隐藏状态 $\mathbf{F}$ 之后添加,以对视频动作进行分类,即

$$p = \text{softmax}(W_F \mathbf{F} + b_F) \quad (20)$$

3 实验结果与分析

3.1 数据集

本文在两个广泛使用的数据集 UT-Interaction (Ryoo等,2010)和 SBU (Yun等,2012)上评估本文方法,并与其他交互识别方法进行比较。UT-Interaction 包含6个交互动作类型,包括握手、拥抱、踢、指、出拳和推,是在两种不同的条件下记录的,因此细分为 UT-1 和 UT-2,共120个视频,每组各一半。每个视频大约有2~6s,帧率为30帧/s。这个数据集上的实验采用10折交叉验证方法。该数据集背景几乎静止,遮挡范围有限,视点固定,分类难度较低。SBU有8种交互类型动作,包括接近、离开、推、踢、打、交接、拥抱和握手。该数据集总共282个短视频(每个大约2~3s)。记录是在相同的实验室环境中进行的,帧率为15帧/s。实验采用的是无噪音数据版本。这个数据集上的实验采用5折交叉验证方法进行评估。

3.2 数据处理

由于 UT-Interaction 只包含 RGB 信息,没有提供明确的姿态信息,因此使用 OpenPose 算法 (Cao等,2017)来提取这类信息,估计视频中对象的关节坐标。默认选项运行,虽然默认选项不是最精确,但对于本文实验来说,速度更快,能够满足精度要求。然而,由于输出是基于帧的,必须采用一些后处理来保证整个视频中参与者1和参与者2身体的一致性。主要根据当前帧与前一帧之间的关节距离,假设连续帧之间的总距离不应该有太大的差异,将每一帧的姿态正确地分配到各自的身体上。对于 UT-Interaction 数据集,将准备好的数据使用 OpenPose 算法提取每个视频每一帧的人体骨架15个关

节点(例如头部、颈部、肩部和手部等)的信息 (x, y, z) , x 为关节在图像上的横坐标, y 为关节在图像上的纵坐标, z 为关节的置信值, 骨架结构图如图 5 所示。对于一个 batch 的视频, 用 4 维矩阵 (N, C, T, V) 来表示。 N 表示 1 个 batch 训练视频的数量, C 代表关节的特征维度, 即 $3(x, y, z)$, T 代表一个视频帧的数量, V 代表关节的数量, 这里是 30 个关节(2 个人关节点数)。对于 SBU 交互数据集, 采用同样的 4 维矩阵 (N, C, T, V) 来表示。

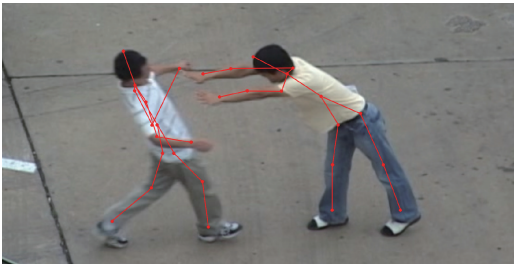


图 5 骨架结构
Fig. 5 Skeleton structure

本文模型基于 PyTorch 框架。在图卷积层中, 将 Chebyshev 阶 K 设为 4。对于切片 RNN 层, 设置隐藏层节点数为 50, 在 UT-Interaction 数据集上切片参数 k 取 2, n 取 5; 在 SBU 数据集上切片参数 k 取 3, n 取 3, 为避免过拟合, dropout 设为 0.5。

3.3 实验结果与分析

在 UT-Interaction 数据集和 SBU 数据集上分别测试, 评估本文算法中各部分的有效性, 为了更好地评估本文算法中各部分的影响, 在 SBU 数据集上, 对不同的连接方法进行实验, 并将本文算法与之前在 UT-Interaction 数据集和 SBU 数据集上的工作进行了比较, 验证了其优越性。

3.3.1 消融实验

为了验证帧内交互建模、帧间建模以及切片 RNN 的有效性, 在 SBU 数据集上进行消融实验, 实验结果如表 1 所示, 将原始的骨架图输入 GCN 的方法作为基线模型 GCN-GRU(Bone only), 该方法获得 93.5% 的精度, 当使用切片 RNN 与 GCN 结合, 模型识别效果并没有得到提升; 当添加帧内交互设计数据特征, 传统 RNN 模型和切片 RNN 模型精度分别提升了 0.7%、0.9%, 验证了帧内交互设计的有效性, 但是切片 RNN 模型仍旧没有比 RNN 模型有着明显的优势; 在帧内交互设计基础上引入帧间设

计数据特征时, 传统 RNN 模型和切片 RNN 模型精度分别为 94.7%、95.2%, 分别提升了 0.5%、1%, 验证了帧间交互设计的有效性, 同时切片 RNN 模型比传统 RNN 模型精度高了 0.5%, 验证了切片方法的有效性。

表 1 消融实验
Table 1 Ablation experiments

算法	准确率/%
GCN-GRU(Bone only)	93.5
GCN-SGRU(Bone only)	93.3
GCN + Intra-connection-GRU	94.2
GCN + Intra-connection-SGRU	94.2
GCN + Intra + Inter-connection-GRU	94.7
GCN + Intra + Inter-connection-SGRU	95.2

3.3.2 对比实验

分别在 UT-Interaction 数据集和 SBU 数据集上进行测试, 并将本文算法获得的动作识别性能与之前的工作进行了比较。本文算法的混淆矩阵如图 6 和图 7 所示, 可以发现在 UT-Interaction 数据集和 SBU 数据集的各个类上都是对角占优的, 这表明本文算法在这两个数据集上取得了很好的分类效果。但是仍然有些行为可能会被贴上错误的标签, 因为它们本身就很相似, 即使是人类的感知也很难区分。本文算法交互连接能很好地提取出对象之间的关系, 减少此类误差的出现。

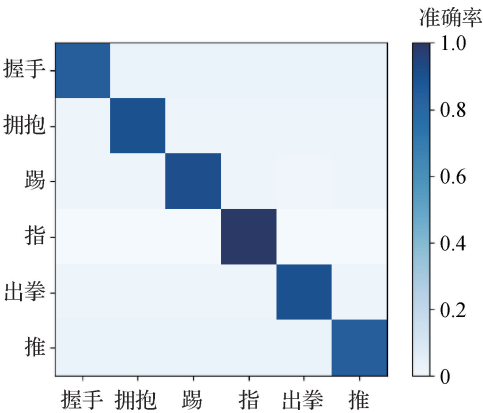


图 6 UT-Interaction 数据集混淆矩阵
Fig. 6 UT-Interaction dataset confusion matrix

在 UT-Interaction 数据集上不同算法结果如表 2 所示。本文算法通过对参与者身体各部分之间的相互作用关系建模, 模型识别精度为 94.6%。实验结

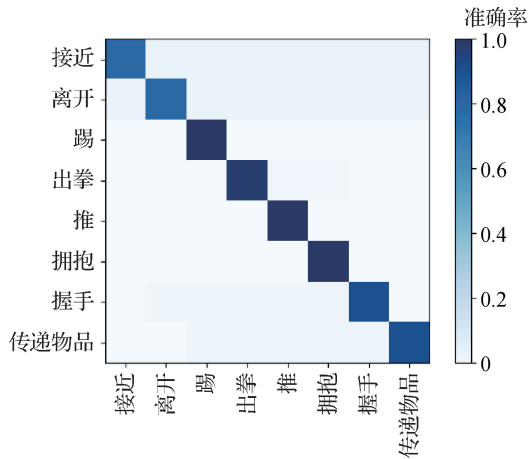


图 7 SBU 数据集混淆矩阵

Fig. 7 SBU dataset confusion matrix

表 2 UT-Interaction 实验效果对比

Table 2 UT-Interaction experimental effect comparison

算法	准确率/%
Yu 等人(2010)	83.33
Raptis 和 Sigal(2013)	93.30
Donahue 等人(2015)	85.00
Aliakbarian 等人(2017)	90.00
H-LSTCM(Shu 等,2021)	98.30
Gao 等人(2019)	93.90
本文	94.60

注:H-LSTCM 为 hierarchical long-short-term concurrent memory。

果表明,该算法可以提高大多数交互的识别结果。同时在实验过程中发现对比实验方法,对于涉及同步动作的交互行为识别,如出拳和推的识别精度较低,由于两者的交互姿势相似,被识别为对方的可能性很大。而本文算法通过交互设计和空间建模的方法能够很好地区分这些相似的动作识别,图 6 所示的混淆矩阵可视化也能够很好地验证,显然本文算法在交互识别方面有较好的效果。Shu 等人(2021)方法以分层的方式从静态的单人特征中学习所有人之间的动态相关表示,并取得了较好的效果。然而,本文算法拥有更少的层,因此需要的参数较少,且使用的骨架数据较小,实验较为简单,同时在运行速度上更具优势。与 Gao 等人(2019)采用谱图卷积结合标准 2D 卷积的方法进行对比,本文算法精度提高了 0.7%,通过切片 RNN 对时序依赖信息的提取,验证了谱图卷积方法与切片 RNN 结合的方法的

有效性。

表 3 给出了不同方法在 SBU 数据集上的比较结果。与其他算法,包括手工制作的基于特征的方法和深度学习方法相比,本文算法取得了最好的性能。在 SBU 数据集中,握手和传递物品出现了一定的误差,本文算法提出的交互连接能够很好地提取出对象之间的关系,减少此类误差的出现。由图 7 可见,由于靠近和离开的动作之间没有交互动作,因此,对于靠近和离开的检测结果较低。但总体上看,本文算法识别效果更具优势,且对于交互类的识别效果更好。

表 3 SBU 实验效果对比

Table 3 SBU experimental effect comparison

算法	准确率/%
Yun 等人(2012)	80.30
Ji 等人(2014)	86.90
STA-LSTM (Song 等,2017)	91.51
GCNConv(Kipf 和 Welling,2017)	90.00
RotClips + MTCNN(Ke 等,2018)	94.17
SGC(Wu 等,2019)	94.00
本文	95.20

注: STA-LSTM 为 spatio-temporal attention long short-term memory。

综上,本文算法能更准确地描述交互作用,对交互作用识别非常有效,特别是对包含同步动作的复杂交互作用识别,时序依赖信息能够提升视频动作识别的精度。

4 结 论

提出了一种基于骨骼的交互动作识别方法,为了有效地描述交互信息,提出了时空建模的方法,将帧内交互建模设计与帧间建模相结合。为了捕获同一对象的非物理连接、不同对象和不同帧之间的连接,充分利用人体关节间的空间和时间依赖性,发现关节间的潜在关系,从而更好地进行识别。对交互行为骨骼图进行有效表示,然后利用谱图卷积和切比雪夫近似进行空间特征提取、使用切片 RNN 对时序依赖信息进行提取。本文算法提高了交互动作识别的准确性,实验表明了该方法的优越性。本文算

法针对的是交互动作的识别,后期会进一步研究、理解交互中各个对象的动作,分析对象之间动作的先后关系,对其进行结构化输出。

参考文献 (References)

- Aliakbarian M S, Saleh F S, Salzmann M, Fernando B, Petersson L and Andersson L. 2017. Encouraging LSTMs to anticipate actions very early//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 280-289 [DOI: 10.1109/ICCV.2017.39]
- Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1302-1310 [DOI: 10.1109/CVPR.2017.143]
- Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T and Saenko K. 2015. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 2625-2634 [DOI: 10.1109/CVPR.2015.7298878]
- Du Y, Wang W and Wang L. 2015. Hierarchical recurrent neural network for skeleton based action recognition//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1110-1118 [DOI: 10.1109/CVPR.2015.7298714]
- Gao X, Hu W, Tang J X, Liu J Y and Guo Z M. 2019. Optimized skeleton-based action recognition via sparsified graph regression//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 601-610 [DOI: 10.1145/3343031.3351170]
- Ji Y L, Ye G and Cheng H. 2014. Interactive body part contrast mining for human interaction recognition//Proceedings of 2014 IEEE International Conference on Multimedia and Expo Workshops. Chengdu, China: IEEE: 1-6 [DOI: 10.1109/ICMEW.2014.6890714]
- Ke Q, Bennamoun M, An S, Boussaid F and Sohel F. 2016. Human interaction prediction using deep temporal features//Proceedings of European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 403-414 [DOI: 10.1007/978-3-319-48881-3_28]
- Ke Q H, Bennamoun M, An S J, Sohel F and Boussaid F. 2018. Learning clip representations for skeleton-based 3D action recognition. IEEE Transactions on Image Processing, 27(6): 2842-2855 [DOI: 10.1109/TIP.2018.2812099]
- Kipf T N and Welling M. 2017. Semi-supervised classification with graph convolutional networks [EB/OL]. [2021-01-06]. <https://arxiv.org/pdf/1609.02907.pdf>
- Li B, Cheng Z H, Xu Z H, Ye W, Lukasiewicz T and Zhang S K. 2019. Long text analysis using sliced recurrent neural networks with breaking point information enrichment//Proceedings of ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 7550-7554 [DOI: 10.1109/ICASSP.2019.8683812]
- Liu J, Shahroudy A, Xu D and Wang G. 2016. Spatio-temporal LSTM with trust gates for 3D human action recognition//Proceedings of European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 816-833 [DOI: 10.1007/978-3-319-46487-9_50]
- Liu M Y, Liu H and Chen C. 2017. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition, 68(3): 346-362 [DOI: 10.1016/j.patcog.2017.02.030]
- Raptis M and Sigal L. 2013. Poselet key-framing: a model for human activity recognition//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 2650-2657 [DOI: 10.1109/CVPR.2013.342]
- Ryoo M S, Chen C C, Aggarwal J K and Roy-Chowdhury A. 2010. An overview of contest on semantic description of human activities (SDHA) 2010//Proceedings of International Conference on Pattern Recognition. Istanbul, Turkey: Springer: 270-285 [DOI: 10.1007/978-3-642-17711-8_28]
- Shi L, Zhang Y F, Cheng J and Lu H Q. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12018-12027 [DOI: 10.1109/CVPR.2019.01230]
- Shu X B, Tang J H, Qi G J, Liu W and Yang J. 2021. Hierarchical long short-term concurrent memory for human interaction recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3): 1110-1118 [DOI: 10.1109/TPAMI.2019.2942030]
- Song S J, Lan C L, Xing J L, Zeng W J and Liu J Y. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press: 4263-4270
- Tang Y S, Tian Y, Lu J W, Li P Y and Zhou J. 2018. Deep progressive reinforcement learning for skeleton-based action recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5323-5332 [DOI: 10.1109/CVPR.2018.00558]
- Van Gemeren C, Tan R T, Poppe R and Veltkamp R C. 2014. Dyadic interaction detection from pose and flow//Proceedings of International Workshop on Human Behavior Understanding. Zurich, Switzerland: Springer: 101-115 [DOI: 10.1007/978-3-319-11839-0_9]
- Wang S G, Sun A M, Zhao W T and Hui X L. 2015. Single and interactive human behavior recognition algorithm based on spatio-temporal interest point. Journal of Jilin University (Engineering and Technol-

ogy Edition), 45(1): 304-308 (王世刚, 孙爱朦, 赵文婷, 惠祥龙. 2015. 基于时空兴趣点的单人行为及交互行为识别. 吉林大学学报(工学版), 45(1): 304-308) [DOI: 10.13229/j.cnki.jdxbgxb201501044]

Wang X H and Deng H M. 2020. A multi-feature representation of skeleton sequences for human interaction recognition. Electronics, 9(1): #187 [DOI: 10.3390/electronics9010187]

Wu F, Zhang T Y, de Souza A H Jr, Fifty C, Yu T and Weinberger K Q. 2019. Simplifying graph convolutional networks [EB/OL]. [2021-01-06]. <https://arxiv.org/pdf/1902.07153.pdf>

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press: 7444-7452

Yu T H, Kim T K and Cipolla R. 2010. Real-time action recognition by spatiotemporal semantic and structural forest//Proceedings of the British Machine Vision Conference. Aberystwyth, UK: BMVA Press: #52 [DOI: 10.5244/C.24.52]

Yun K, Honorio J, Chattopadhyay D, Berg T L and Samaras D. 2012. Two-person interaction detection using body-pose features and multiple instance learning//Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, USA: IEEE: 28-35 [DOI: 10.1109/CVPRW.2012.6239234]

作者简介



成科扬, 1982 年生, 男, 教授, 主要研究方向为计算机视觉、模式识别。

E-mail: kycheng@ujs.edu.cn



吴金霞, 通信作者, 女, 硕士研究生, 主要研究方向为计算机视觉、模式识别。

E-mail: 2061383373@qq.com

王文杉, 女, 副研究员, 主要研究方向为统计分析、机器学习。

E-mail: wenshanwang2468@163.com

荣兰, 女, 硕士研究生, 主要研究方向为计算机视觉、模式识别。E-mail: 1748675160@qq.com

詹永照, 男, 教授, 博士生导师, 主要研究方向为多媒体、人工智能。E-mail: yzzhan@ujs.edu.cn