智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

深度学习的双人交互行为识别与预测算法研究

姬晓飞， 谢旋， 任艳

引用本文:
姬晓飞, 谢旋, 任艳. 深度学习的双人交互行为识别与预测算法研究[J]. 智能系统学报, 2020, 15(3): 484–490.
JI Xiaofei, XIE Xuan, REN Yan. Human interaction recognition and prediction algorithm based on deep learning[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(3): 484–490.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201812029

DOI: 10.11992/tis.201812029

深度学习的双人交互行为识别与预测算法研究

姬晓飞，谢旋，任艳
（沈阳航空航天大学 自动化学院，辽宁 沈阳 110136）

摘　要：基于卷积神经网络的双人交互行为识别算法存在提取的深度特征无法有效表征交互行为序列特性的问题，本文将长短期记忆网络与卷积神经网络模型相结合，提出了一种基于深度学习的双人交互行为识别与预测一体化方法。该方法在训练过程中，完成对卷积神经网络和长短期记忆网络模型的参数训练。在识别与预测过程中，将不同时间比例长度的未知动作类别的视频图像分别送入已经训练好的卷积神经网络模型提取深度特征，再将卷积神经网络提取的深度特征送入长短期记忆网络模型完成对双人交互行为的识别与预测。在国际公开的 UT-interaction 双人交互行为数据库进行测试的结果表明，该方法在保证计算量适当的同时对交互行为的正确识别率达到了 92.31%，并且也可完成对未知动作的初步预测。

关键词：视频分析；行为识别；行为预测；深度学习；卷积神经网络；长短期记忆网络；UT-interaction 数据库；SBU Kinect interaction 数据库
中图分类号：TP391.4　　**文献标志码：**A　　**文章编号：**1673−4785(2020)03−0484−07

中文引用格式：姬晓飞，谢旋，任艳. 深度学习的双人交互行为识别与预测算法研究 [J]. 智能系统学报，2020，15(3)：484–490.
英文引用格式：JI Xiaofei, XIE Xuan, REN Yan. Human interaction recognition and prediction algorithm based on deep learning[J]. CAAI transactions on intelligent systems, 2020, 15(3): 484–490.

Human interaction recognition and prediction algorithm based on deep learning

JI Xiaofei, XIE Xuan, REN Yan
(School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: A drawback of the human interaction recognition algorithm based on a convolutional neural network (CNN) is that the extracted depth features cannot effectively represent the characteristics of interaction sequences. Instead, this paper proposes a human interaction recognition and prediction algorithm based on deep learning, by combining the Long Short-Term Memory (LSTM) network with the CNN model. In the process, video images of unknown action categories of different time lengths are sent to a trained CNN model to extract the depth features. The depth features are then sent to a trained LSTM model to complete the recognition and prediction of the interaction behavior. When tested on the UT-interaction human interaction behavior dataset, the algorithm demonstrates a 92.31% correct human interaction recognition rate and can complete the preliminary prediction of unknown actions.

Keywords: video analysis; action recognition; action prediction; deep learning; convolutional neural network; long short term memory; UT-interaction dataset; SBU Kinect interaction dataset

基于视频的双人交互行为识别与预测研究备受计算机视觉领域的关注。近期成果有，文献 [1] 率先提出动态 BoW(bag of word) 的概率统计方法解决双人交互行为的预测问题，采用整体直方图对动作的时空特征进行表示，而后建模整体直方图特征随时间变化的规律实现动作预测。文献 [2] 提出一种基于 3D 立体体积局部兴趣点的时空描述符结合稀疏特征表示的预测框架，将每个视频划分为多个时间段，通过构造各个时间段组合稀

收稿日期：2018−12−26.
基金项目：国家自然科学基金项目 (61602321)；辽宁省自然科学基金项目 (201602557)；辽宁省教育厅科学研究服务地方项目 (L201708)；辽宁省教育厅科学研究青年项目 (L201745).
通信作者：姬晓飞. E-mail：jixiaofei7804@126.com.

疏词袋, 判断人类活动类别。文献 [3] 提出了一个新的基于关键帧的动作预测模型, 提取运动视频的关键帧作为运动模型的状态节点, 成功地实现了双人交互行为识别与预测的统一框架。但此方法需要根据关键帧之间的相关性对其编码, 计算量很大且分类模型复杂。文献 [4] 提出一种新的判别式多尺度核化模型, 分别采用局部进度模型和全局进度模型捕获时间进度和全局观测之间的关系, 实现对部分观测视频的动作预测。此方法需要进行复杂的数学建模和时空匹配核函数的选择。基于传统特征提取方法的双人交互行为识别与预测研究中, 特征提取完全是基于经验的手动选择, 适应性不强。基于模板匹配的识别方法[5-7]往往不考虑视频序列的时序关系。基于时序建模的识别方法[8-12]不适用于高维特征的建模与识别。总的来说基于传统特征提取方法的双人交互动作识别与预测的准确率不高, 且提升的空间不大。

目前, 深度学习理论与算法研究已经取得了重大进展, 利用深度学习模型提取特征可以消除人工设计过程中的盲目性和差异性, 实现特征的自动提取。文献 [13] 将深度学习网络用于双人交互行为识别, 首先利用深度卷积神经网络从视频的连续光流图像中提取深度特征, 然后将深度特征按照时间维度连接在一起进行卷积以学习时间信息, 最后采用多层神经网络输出交互类别以实现动作的识别与预测。该算法侧重于时间特征提取和建模, 忽略了空间信息对于识别结果的影响。文献 [14] 采用空间和时间卷积神经网络, 提出从动作视频中学习空间和时间信息的双流方法, 然后用平均时间和空间两个流的输出概率分数来识别视频中动作类别。这种双流方法得到了较高的识别率, 但其忽略了交互场景上下文中的重要序列信息。此外上述两种方法输入均为视频的光流信息, 计算量较大难以实现实时操作。根据以上分析, 为了提取适应性较强的特征表示, 本文选取卷积神经网络提取动作视频深展特征, 考虑到单纯使用卷积神经网络提取的深度特征无法准确有效表征交互行为序列特性的问题, 将长短期记忆网络 (long short term memory network, LSTM)[15] 与卷积神经网络 (convolutional neural networks, CNN) 模型相结合, 提出了一种基于深度学习的双人交互行为识别与预测一体化方法。该方法充分利用 CNN 和 LSTM 的优势来提取和建模两个相互作用的个体之间的长期相互关联特性, 提高了交互行为识别与预测准确率。

1　基于深度学习的算法框架

本文算法的处理框架如图 1 所示。

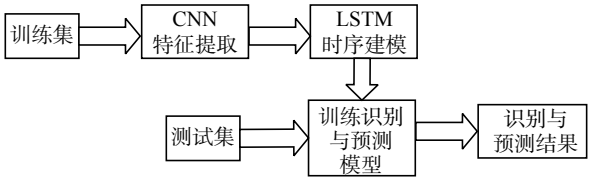


图 1　算法处理框架

Fig. 1　Algorithm flowchart

本文算法处理流程为:

1) 在训练过程中, 完成对 CNN 和 LSTM 模型的参数训练, 即将训练视频中的所有帧图像分别送入卷积神经网络中提取深度特征, 然后卷积神经网络全连接层输出结果作为 LSTM 的输入。

2) 在识别与预测过程中, 直接将不同时间比例长度的未知动作类别的测试视频帧图像分别送入已经训练好的识别与预测模型, 最终得到每类动作的检测分数, 从而实现对双人交互行为的识别与预测。

该算法的优势在于利用卷积神经网络提取鲁棒性极强的深度特征, LSTM 完成视频帧图像的时序建模, 充分考虑到时间和视频上下文信息。

2　视频预处理

双人交互行为涉及手或腿等四肢的伸展, 这些身体部位对于准确地进行交互行为识别与预测非常重要。但是, 人类的边界框并不完全包括所有身体部位的伸展。在这种情况下, 为了提高交互行为识别与预测的准确性, 在特征提取之前通过帧间差分的方法获得图像剪影信息, 然后合并两个人的边界框来选择感兴趣区域 (ROI), 最后通过裁剪 ROI 对每个输入图像进行归一化, 如图 2 所示。

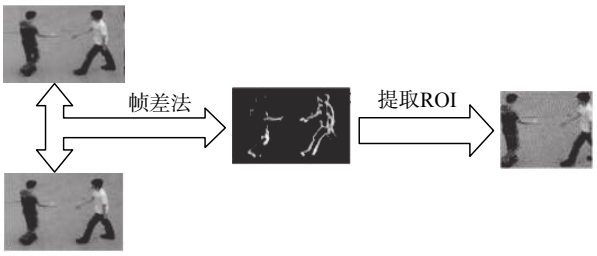


图 2　帧差法提取剪影所在感兴趣区域

Fig. 2　Frame difference method extracts the ROI

3　特征表示

3.1　卷积神经网络概述

CNN 是一种典型的深度学习网络[16], 由卷积

层、池化层和全连接层构成。基本结构框图如图 3 所示。

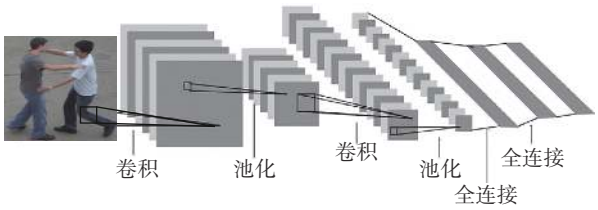


图 3　CNN 基本结构框图

Fig. 3　CNN basic structure block diagram

1) 卷积层 (convolutional layer): 卷积层可以提取输入数据的特征，每个卷积层只能提取一些简单特征，但是包含很多卷积层的深度神经网络可以抽取复杂的抽象特征表示。每个卷积单元的参数都是通过反向传播算法优化计算得到的。

2) 池化层 (pooling layer): 也叫作降采样层。数据经过卷积层处理之后会得到维度非常大的特征，池化层将这些特征分割成几个区域，取其最大值或平均值，得到新的、维度较小的特征。

3) 全连接层 (fully-connected layer): 把所有局部特征组合成全局特征，用来计算每类动作的识别概率或者将深度特征输入到后续用于时序建模的深度网络中。

3.2　InceptionV3 网络

本文采用 GoogLeNet 网络[17] 的改进版本 InceptionV3[18] 深度卷积神经网络架构。GoogLeNet 对网络中的传统卷积层进行了修改，主要特点在保证没有增加计算量的前提下，提高网络内部计算资源的利用率，允许增加网络深度和广度来提高深度神经网络的性能。该 GoogLeNet 模型近两年来经过一系列网络模型与参数的改进，网络的最后全连接层的输出用作输入图像的深度特征表示。经过大量实验验证，采用 InceptionV3 得到的预测与识别效果最佳，Inception V3 中 Inception 模块重复很多次，最终构成了 GoogLeNet 22 层的深层模型，输出 2048 维深度特征向量。

4　时序建模

4.1　LSTM 网络

传统的神经网络没有记忆功能，模型不关注上一时刻处理信息会有哪些用于下一时刻，每一次都只会关注当前时刻的处理过程，所以在双人交互行为的序列学习中，单纯地采用传统神经网络方法往往存在忽略时间信息或无法考虑上下文关联信息的问题。递归神经网络 (recurrent neural network, RNN) 结构带有一个指向自身的环，用来表示可以将当前时刻处理的信息传递给下一时刻使用。但传统的 RNN 记忆能力较弱，只能学习短时间内上下文信息，而 LSTM 是传统 RNN 的改进网络，主要用于改善传统 RNN 的弱记忆能力限制的问题。

LSTM 网络结构如图 4 所示，图中给出了相同的单元如何在生成输出流 h 的同时响应输入流 x，其中每个单元由 3 个门组成，即忘记门、输入门和输出门。该单元设计为从最初观测时刻到当前时刻进行记忆，并且在记忆过程中会丢弃不需要的存储单元。输入门用来确定需要处理的信息，输出门用来选择输出通道。

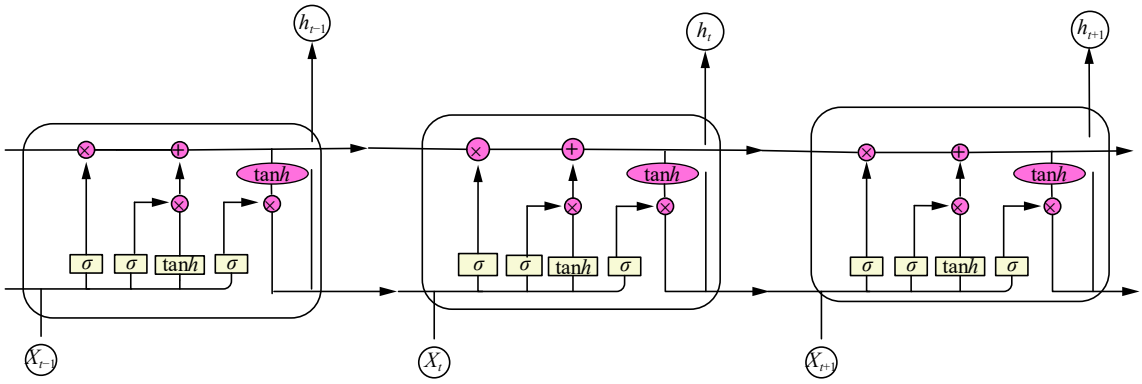


图 4　LSTM 结构图

Fig. 4　LSTM structure block diagram

在数学上，LSTM 存储器单元在给定时刻 t 的瞬时输出 h_t 定义为

$$h_t = f(x_t, h_{t-1}) \tag{1}$$

式中：x_t 是当前输入量；h_{t-1} 是与前一时刻关联的记忆响应，f 是 LSTM 单元学习的非线性函数映射。本质上，LSTM 在当前时刻的输出是从初始

时刻开始一直到上一时刻响应的递归函数。

与传统 RNN 相比，LSTM 网络是具有记忆单元的 RNN，包含能够学习长期依赖性的存储单元，不仅可以用于捕获和存储前观察信息，并且可以提供更长范围的上下文信息进行当前的预测。

4.2　LSTM 和 CNN 模型结合

LSTM 在提取序列的时序特征上效果较好，适用于人体动作的识别与预测问题的研究。因此，本文将卷积神经网络提取的深度特征送到 LSTM 进行时序建模，以充分捕捉视频特征的时间和上下文信息。该网络模型将输入视频帧图像送入 CNN 结构进行特征变换产生固定长度的特征向量表示 (本文深度特征输出向量为 2048 维)。然后将 CNN 的输出输送到时间序列学习模块即一系列 LSTM 单元中，最终对所有帧的概率分布进行平均，选择最有可能的标签，完成视频动作的分类，如图 5 所示。

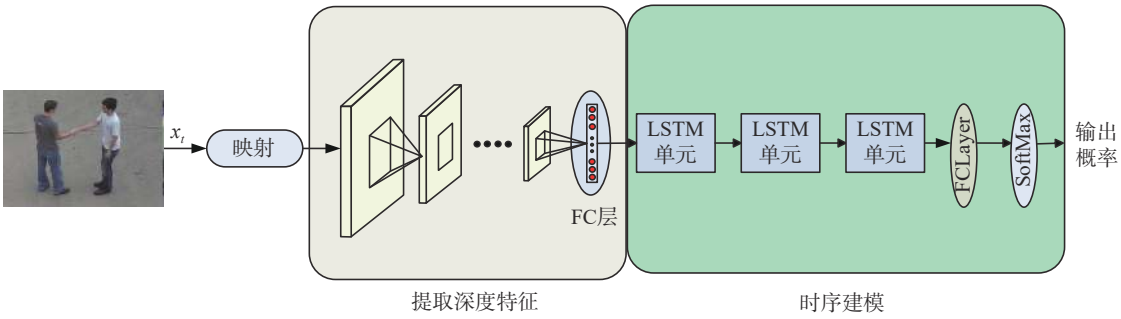


图 5　单帧图像处理流程

Fig. 5　Single frame image processing flow

5　实验结果与分析

5.1　数据库信息

本文提出了一种基于深度学习的双人交互行为识别与预测算法，为了充分证明算法的有效性和合理性，在公开的 UT-interaction[19] 双人交互行为数据库 set1 数据集中进行测试。该数据集包括握手、拥抱、踢打、指、拳击和推搡 6 类动作。此数据库每类动作由不同的人来完成，共有 60 组交互动作。库内动作没有周期性规律，并且不同动作类别之间存在相似性动作，所以对库内交互行为动作识别与预测更具有挑战性。数据集包含的动作如图 6 所示。

图 6　UT-interaction 数据库示例

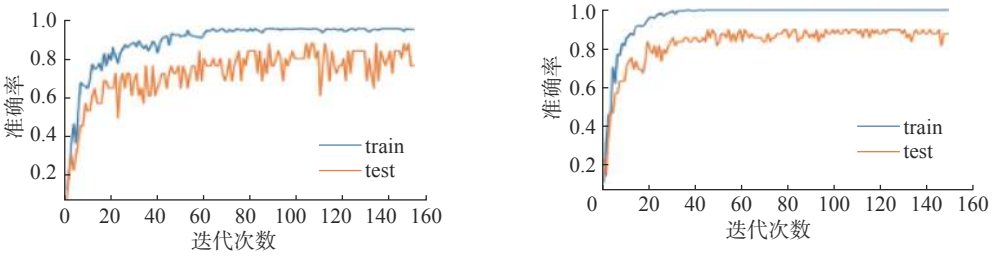
Fig. 6　Exemplar frames from UT-interaction dataset

5.2　实验测试结果

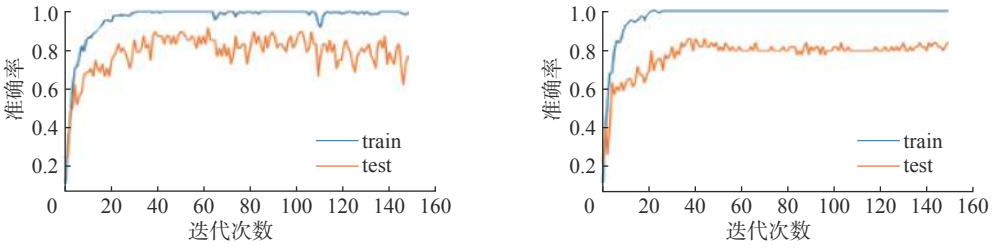
本文首先对视频帧图像进行预处理，去掉冗余的干扰信息。然后将处理过的视频图像送到 InceptionV3 网络模型中提取深度特征，完全连接层 FC 输出 2048 维深度特征向量，送入 LSTM 网络模型中作为其输入量完成对输入视频的连续多帧图像的时序建模，最终输出当前测试视频的动作类别分数，从而判断未知测试视频的动作类别。所有实验在主频为 3.50 GHz，内存 32 GB，带有两个显卡 NVIDIA1080TI 的 64 位 Ubuntun16.04 LTS 操作系统下完成，处理器为 i7-7800X，实验软件平台为 python3.6.4 keras2.1.3。

本次实验的结果如图 7 和图 8 所示。由图 7 可知，随着测试视频时间长度的增加，预测准确率逐渐增加。当测试视频长度比例为 100% 时，双人交互行为预测问题退化为识别问题；当测试视频长度比例为 50% 时，预测准确率达到 78.85%，

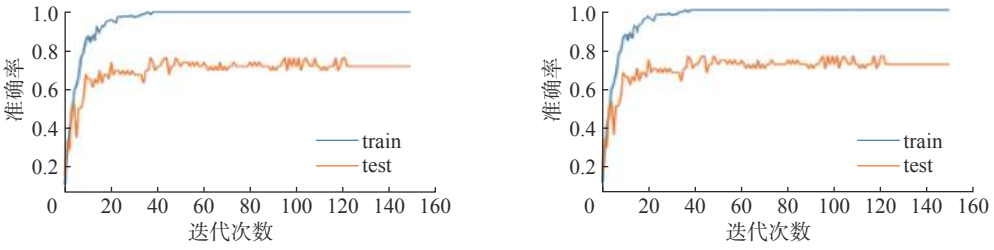
进一步验证了本文提出的基于深度学习的方法可以完成交互行为的识别与预测一体化。由图 8 可知，随着训练次数增多，模型学习能力不断增强，对数据的拟合程度也不断提高。



(a) 测试视频为100%，识别准确率为92.31%　(b) 测试视频为90%，预测准确率为90.38%



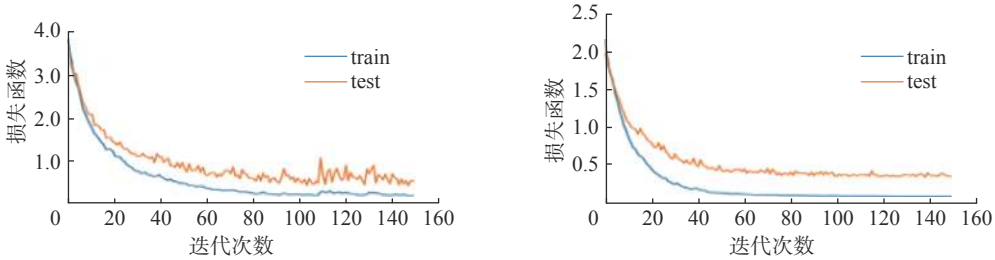
(c) 测试视频为80%，预测准确率为90.00%　(d) 测试视频为70%，预测准确率为86.54%



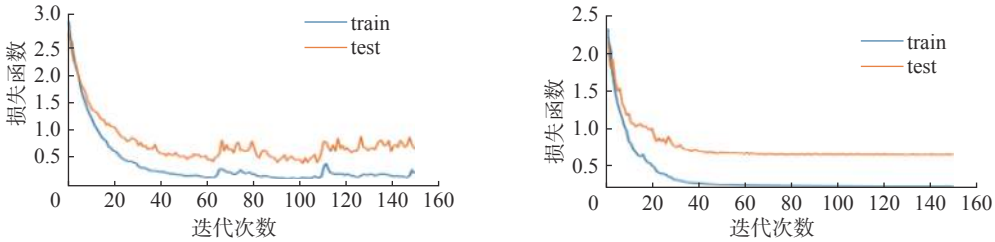
(e) 测试视频为60%，预测准确率为79.00%　(f) 测试视频为50%，预测准确率为78.85%

图 7　不同时间长度下预测准确率

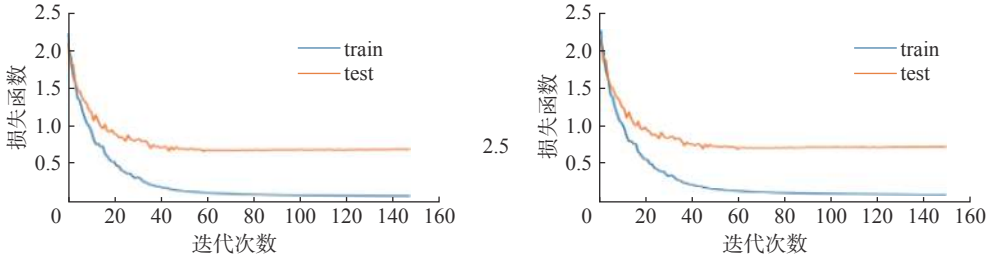
Fig. 7　Prediction accuracy at different lengths of time



(a) 测试视频时间长度为100%　(b) 测试视频时间长度为90%



(c) 测试视频时间长度为80%　(d) 测试视频时间长度为70%



(e) 测试视频时间长度为60%　　(f) 测试视频时间长度为50%

图 8　不同时间长度下模型损失

Fig. 8　Model loss at different lengths of time

5.3　实验结果与其他文献的比较

本文将提出的方法与其他文献在 UT-interaction 数据库中得到的识别与预测结果进行比较，见表 1 所示。

从表 1 可以看出，本文提出的基于深度学习的新框架在双人交互行为识别与预测一体化上得到了较好的结果。与文献 [1-4] 相比，本文的方法在对未知动作类别视频的预测问题上准确性最优，尽管文献 [3] 和文献 [4] 得到的识别结果稍高于本文的算法，但是其算法较为复杂，在前期对输入图像的预处理计算量很大。文献 [3] 实验结果依附于带有明显判别动作的关键帧信息，算法执行性和实际可操作性不强；文献 [4] 开发了一种具有复合内核的最大边际动作预测机制，但是其基于丰富的预测先验知识，学习过程复杂，且很难实现实时操作。本文采用的方法无需人为参与，即可达到较好的识别与预测效果，且可以实现实时操作。但由于深度学习算法对数据量的要求较高，要想达到更优的预测与识别效果，训练模型仍然有待一步进行改善。

表 1　不同方法在 UT-interaction 数据库中的动作识别与预测结果比较

Table 1　Comparison of different algorithms for interaction recognition and prediction in UT-interaction dataset

来源	一半观测序列得到的识别率/%	整个观测序列得到的识别率/%
文献[1]	65.00	81.70
文献[2]	70.00	80.00
文献[3]	73.30	93.30
文献[4]	78.33	95.00
本文	78.85	92.31

6　结束语

本文提出一种基于深度学习的人体交互行为识别与预测方法，来解决基于传统特征的交互行为识别与预测技术复杂、准确性较低的问题。人类交互动作可能持续很长时间，并且可以由多个不同的子动作组成。单纯使用在交互发生之前捕获的单个帧来推断交互类别信息往往是不够的，几个连续帧的时间信息和上下文依赖为预测未来的交互动作提供了关键线索。采用新颖的网络组合模型 LSTM+InceptionV3，旨在了解视频全局和局部上下文之间的依赖关系，并捕获交互场景的显著信息。实验测试结果表明，本文采用的算法在国际公开数据库中取得了良好的结果，识别准确率和算法鲁棒性都有了明显的提升。

参考文献：

[1] RYOO M S. Human activity prediction: Early recognition of ongoing activities from streaming videos[C]//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain, 2011: 1036−1043.

[2] XU Kaiping, QIN Zheng, WANG Guolong. Human activities prediction by learning combinatorial sparse representations[C]//Proceedings of 2016 IEEE International Conference on Image Processing. Phoenix, USA, 2016: 724−728.

[3] RAPTIS M, SIGAL L. Poselet key-framing: a model for human activity recognition[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2650−2657.

[4] KONG Yu, FU Yun. Max-margin action prediction machine[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(9): 1844–1858.

[5] KUNZE K, LUKOWICZ P. Dealing with sensor displacement in motion-based onbody activity recognition systems[C]//Proceedings of the 10th International Conference on Ubiquitous Computing. Seoul, South Korea, 2008: 20−29.

[6] BULLING A, ROGGEN D. Recognition of visual memory recall processes using eye movement analysis[C]//Proceedings of the 13th International Conference on Ubiquitous Computing. New York, USA, 2011: 455−464.

[7] VAN KASTEREN T, NOULAS A, ENGLEBIENNE G, et

al. Accurate activity recognition in a home setting[C]//Proceedings of the 10th International Conference on Ubiquitous Computing. Seoul, South Korea, 2008: 1−9.

[8] CHUNG P C, LIU C D. A daily behavior enabled hidden Markov model for human behavior understanding[J]. Pattern recognition, 2008, 41(5): 1572−1580.

[9] TANG K, LI Feifei, KOLLER D. Learning latent temporal structure for complex event detection[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1025−1257.

[10] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, USA, 2001: 282−289.

[11] ZHANG Jianguo, GONG Shaogang. Action categorization with modified hidden conditional random field[J]. Pattern recognition, 2010, 43(1): 197−203.

[12] SONG Yale, MORENCY L P, DAVIS R. Action recognition by hierarchical sequence summarization[C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3563−3569.

[13] KE Qiuhong, BENNAMOUN M, AN Senjian, et al. Human interaction prediction using deep temporal features [C]//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 403−414.

[14] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 568−576.

[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735−1780.

[16] BACCOUCHE M, MAMALET F, WOLF C, et al. Sequential deep learning for human action recognition[C]// Proceedings of the 2nd International Workshop on Human Behavior Understanding. Amsterdam, The Netherlands, 2011: 29−39.

[17] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1−9.

[18] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2818−2826.

[19] RYOO M S, AGGARWAL J K. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities[C]//Proceedings of 2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan, 2009: 1593−1600.

作者简介：

姬晓飞，副教授，博士，主要研究方向为视频分析与处理、模式识别理论。承担国家自然科学基金、辽宁省自然科学基金等多项课题研究。发表学术论文 40 余篇，参与编著英文专著 2 部。

谢旋，硕士研究生，主要研究方向为生物特征识别与行为分析技术。

任艳，讲师，博士，主要研究方向为基于公理化模糊集的知识发现与表示、图像语义特征提取。承担国家自然科学基金、航空基金、辽宁省自然科学基金等课题研究。发表学术论文 25 篇。