DOI: 10.13382/j.jemi.B2104116

CNN A-BLSTM network 的双人交互行为识别*

赵 挺[1] 曹江涛[1] 姬晓飞[2]
(1. 辽宁石油化工大学信息与控制工程学院 抚顺 113001; 2. 沈阳航空航天大学自动化学院 沈阳 110136)

摘 要:关节点数据结合卷积神经网络用于双人交互行为识别存在图像化过程中对交互信息表达不充分且不能有效建模时序关系问题,而结合循环神经网络中存在侧重于对时间信息的表示却忽略了双人交互空间结构信息构建的问题。为此提出一种新的卷积神经网络结合加入注意机制的双向长短时期记忆网络(CNN A-BLSTM)模型。首先对每个人的关节点采用基于遍历树结构进行排列,然后对视频中的每一帧数据构建交互矩阵,矩阵的中的数值为排列后双人之间所有的关节点坐标间的欧氏距离,将矩阵进行灰度图像编码后所得图像依次送入 CNN 中提取深层次特征得到特征序列,然后将所得序列送入 A-BLSTM 网络中进行时序建模,最后送入 Softmax 分类器得到识别结果。将新模型用于 NTU RGB D 数据集中的 11 类双人交互行为的识别,其准确率为 90%,高于目前的双人交互行为识别算法,验证了该模型的有效性和良好的泛化性能。

关键词: 双人交互行为识别;深度学习;卷积神经网络;双向长短时期记忆网络;注意机制
中图分类号: TP391.41 **文献标识码**: A **国家标准学科分类代码**: 510.40

CNN A-BLSTM network for two-person interaction behavior recognition

Zhao Ting[1] Cao Jiangtao[1] Ji Xiaofei[2]
(1. School of Information and Control Engineering, Liaoning Petrochemical University, Fushun 113001, China;
2. School of Automation, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: Joint data combined with convolutional neural network for two-person interaction behavior recognition has the problem of insufficient expression of interactive information during the imaging process and ineffective modeling of time-series relations. In combination with recurrent neural network, there is a problem that focuses on the representation of time information. However, it ignores the problem of constructing information about the spatial structure of the two-person interaction. Therefore, a novel model named CNN attention-bidirectional long short-term memory (CNN A-BLSTM) network is proposed. First, the joints of each person are arranged based on the traversal tree structure, and then the interaction matrix is constructed for each frame of data in the video. The values in the matrix are the Euclidean distance between the arranged joint coordinates of two persons. After encoding the gray-scale image of the matrix, the images are sequentially sent to CNN to extract deep-level features to obtain the feature sequence. And then the obtained feature sequence is sent to the A-BLSTM network for time series modeling, and finally sent to the Softmax classifier to obtain the recognition result. The new model is applied to 11 types of two-person interaction in NTU RGB D dataset, and the accuracy is 90%, which is higher than the current two-person interaction recognition algorithm. The effectiveness and good generalization performance of the new model are verified.

Keywords: two-person interaction behavior recognition; deep learning; convolutional neural network; bidirectional long short-term memory network; attention mechanism

0 引 言

视频中的双人交互行为识别是计算机视觉的一个重要研究领域,已广泛地应用于视频理解、分类以及智能监控等场景[1-2]。用于行为识别的主要数据源有 RGB[3]、深度视频[4]和关节点数据[5]。随着高性价比深度相机的普及,获取人体关节点数据的成本大大降低,同时提高了

收稿日期: 2021-04-01 Received Date: 2021-04-01
*基金项目:国家自然科学基金(61673199)、辽宁省科技公益研究基金(2016002006)项目资助

全身关节点数据的跟踪精度。与 RGB 和深度视频相比，关节点数据包含了人体主要关节的三维位置，对视角、身体尺度、运动速度、光照等外部环境的变化具有较强的鲁棒性[6]。因此，基于关节点数据的双人交互行为识别受到越来越多的关注[7-8]。

目前基于关节点数据的双人交互行为识别研究方法主要分为传统方法与深度学习（deeping learning，DL）的方法[9]。传统方法从关节数据中构造并提取特征后再送入分类器中进行双人交互行为识别。Yun 等[10]利用两个人所有关节点之间的距离，关节点运动距离，关节点与平面之间的距离，以及速度特征来表示交互行为，然后送入支持向量机（support vector machine，SVM）进行识别。Vemulapalli 等[11]将骨架结构和动作分别表示为 Lie 群中的点和曲线，并使用 SVM 分类器对动作进行分类。Ji 等[12]在运动描述的短帧集合中计算了八对交互肢体的时空关节姿态特征，构建一个姿态词典对交互行为进行表示，再采用 SVM 进行分类识别。传统方法需要调节大量的参数来提取具有较高辨识力的特征，且均在小数据集上进行测试，模型的迁移性差，识别率很难进一步提高。因此本文拟采用深度学习的方法自动提取判别性特征，增强模型的泛化能力。

基于深度学习的方法主要以卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural networks，RNN）为基础网络展开研究。基于卷积神经网络的方法主要将关节点的位置或运动轨迹进行时空编码成图像后送入 CNN 提取深层次信息进行识别。Ke 等[13]提出一种学习片段表达的方法，生成关节点三维坐标（x，y，z）的 3 个通道分别对应的 3 个片段后再送入 CNN 中进行识别。Caetano 等[14]构造了一种新的树状结构参考关节图像（tree structure reference joints image，TSRJI）分别送入 CNN 中并采用后期融合进行识别。与传统方法相比，基于 CNN 的方法的识别准确率得到了大幅度地提升。然而，基于 CNN 的方法依赖于从关节点数据及其特征编码图像的表现能力，在这一过程中，仅仅使用简单的行或列在时间轴的堆叠来表示时间信息，不能很好地建模视频的时序关系，且大部算法没有很好的表征双人交互的关系。因此，本文拟引入时序建模网络以有效地对视频的时序信息进行建模。

基于循环神经网络的方法主要利用原始关节点数据作为输入，可以有效地对时间信息进行建模。Liu 等[15]提出了一种时空长短期记忆（spatial-temporal long short-term memory，ST-LSTM）网络，将传统的基于 LSTM 的学习扩展到时空域。通过对原始关节数据进行树状结构排列建立关节点之间的依赖性，最后送入 LSTM 中进行建模与识别。Lee 等[16]提出了一种包含短期、中期和长期的集成时间滑动的长短时期记忆网络（temporal sliding LSTM，TS-LSTM）进行行为识别。Li 等[17]提出了一种新的独立循环神经网络（independent recurrent neural network，IndRNN）用于双人交互行为识别，在该网络中，同一层的神经元相互独立，在不同的层上相互连接，可以有效地防止梯度爆炸和消失，使网络能够学习长期依赖性。基于 RNN 的方法往往侧重于对时间信息的表示[18]，而忽略了人体运动中的空间结构信息及其对应的语义信息和两个人之间的交互关系的构建。因此，与上述文献直接采用关节点数据作为输入不同，本文在底层通过计算双人关节点之间的欧氏距离加强空间信息以及交互关系的表达。

综上所述，关节点的空间信息的描述和视频帧间的时间信息的建模是基于关节点数据的双人交互行为识别的最重要的两个因素，时空信息结合是最有效的表示方法。因此，针对双人交互的底层特征设计，本文提出一种新的双人交互底层特征图像化表示，即交互矩阵图，不仅可以更加有效地恢复单帧图像的空间信息，而且可以很好的对交互个体的交互关系进行表示。并进一步提出了一种新的 CNN 结合加入注意机制的双向长短期记忆网络（CNN attention-bidirectional LSTM，CNN A-BLSTM）的双人交互行为识别算法，该算法框架如图 1 所示，首先对原始双人关节点序列进行树形结构的重新排列，并为视频中的每一帧数据构造交互矩阵。矩阵的行和列分别表示第 1 个人和第 2 个人通过计算所得到的关节点坐标之间的欧氏距离，然后将其归一量化为灰度图像得到交互矩阵图序列，以加入双人之间重要的交互信息。然后利用 CNN（ResNet50）提取每一帧交互矩阵图的深层次特征，以增加特征之间的判别性。将所得特征序列送入 A-BLSTM 进行关键帧筛选并建模时序关系，以关注于具有较强辨识力的帧，排除与识别的行为相关性较低的噪声帧的影响，恢复视频的时序信息。最后送入 softmax 分类器中得到最终的识别结果。

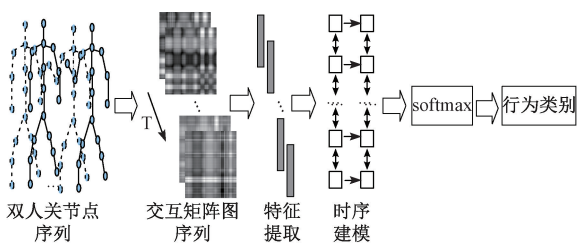


图 1　算法框架

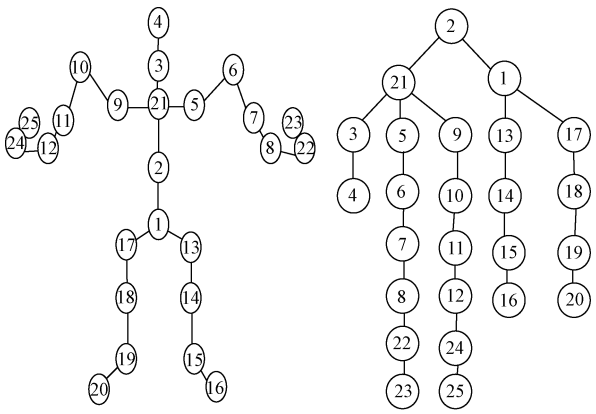
Fig. 1　Algorithm framework

1　关节点数据图像化

目前，CNN 只能对图像数据学习判别性特征，对于非图像数据，需要通过有效的方式进行图像化后才能送入网络中进行学习。因此，将关节点数据送入 CNN 中提取深层特征之前需要将其进行图像化，保留合理且丰富

的空间结构信息以及交互个体之间的交互关系。

采集到的关节点数据形式为给定三维关节点坐标 $J=(x,y,z)$，一个人的关节点表示为一组关节坐标 $S=\{J_1,J_2,\cdots,J_N\}$，其中 N 为每个人的关节数。大部分的研究在将关节数据进行图像化的过程中只是将关节按照固定顺序的链接进行排列，如图 2(a)所示，从而忽略了关节之间的依赖性和更好的空间关系。因此，受文献[15]的启发，对关节点进行树结构排列，如图 2(b)所示。其遍历树顺序为[2, 21, 3, 4, 3, 21, 5, 6, 7, 8, 22, 23, 22, 8, 7, 6, 5, 21, 9, 10, 11, 12, 24, 25, 24, 12, 11, 10, 9, 21, 2, 1, 13, 14, 15, 16, 15, 14, 13, 1, 17, 18, 19, 20, 19, 18, 17, 1, 2]。该结构可以更好地建立个体自身关节之间的依赖性，更加符合人体关节运动趋势，从而获得更加丰富的空间结构关系。



(a) 关节点结构顺序　(b) 关节点树结构
(a) Structure order of joints　(b) Joint tree structure

图 2　所应用关节点数据示例

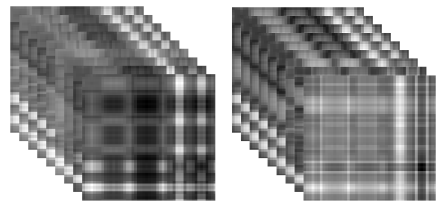
Fig. 2　Examples of applied joint data

直接对关节点数据进行顺序排列图像化方法存在抗干扰性差，对应的语义不明确，结构信息丢失以及交互关系表述不充分等问题，因此提出在关节点树形结构的基础上，构建双人交互距离矩阵，对双人姿态信息和双人的交互信息进行图像化表示，即矩阵的行和列分别代表第 1 个人与第 2 个人经过遍历树顺序重构后的 49 个关节点间的欧氏距离。计算如下：

$$D_{ij}=\| p_i^1-p_j^2 \| \tag{1}$$

式中：p_i^1 和 p_j^2 表示为第 1 个人和第 2 个人重构后的关节点坐标；D_{ij} 为关节点间的距离。最后，将单帧得到的交互距离矩阵归一化到 0~255，得到一张 49×49 的灰度图，对一个视频里所有帧执行该操作则得到交互矩阵图序列。交互矩阵图在增强个体自身关节点之间的依赖性的同时，还考虑了双人之间随时间变化的关节点间的距离特征，可以更加有效地表达双人交互行为的空间信息。

如图 3 所示，交互矩阵图包含大量的信息，但是这些图中的图像特征并不明显，对其分析存在直接识别比较困难的问题，因此将其送入 CNN 中提取深层次的语义特征，增强特征的判别性，有助于后续的建模与识别。



(a) 靠近　(b) 远离
(a) Walking towards　(b) Walking apart

图 3　图像化结果示例

Fig. 3　Example of imaging results

2　特征提取与时序建模

2.1　特征提取

CNN 是一种多层神经网络模型，通过多次卷积和池化运算，可以将大量数据降维，挖掘图像的深层次特征，在图像分类任务上有着良好的表现。目前著名的卷积神经网络有 AlexNet、VGG、Inception 和 ResNet。但在很多情况下，随着模型深度的增加，模型中权重的梯度会越来越小，或者相反地在每一层中爆炸式地增大，从而阻止模型的改进。如图 4 所示，加入残差模块的 ResNet 架构已被证明是解决深度神经网络进行反向传播时常见的梯度消失或爆炸问题的可靠方案[19]。因此，本文采用 ResNet-50 通过迁移学习[20]的方式作为特征提取的基本模块。具体方法为将一个视频得到的每一帧大小为 的交互矩阵图缩放为，以适应 ResNet-50 网络的图像输入尺寸，只采用卷积层对图像进行特征提取，得到每一帧维数为 2 048 的特征向量序列，然后将得到的特征序列送入后续的 A-BLSTM 网络中进行时序建模。

2.2　时序建模

1)注意机制

视觉注意机制在很多领域都取得了成功，包括基于 RGB 视频的行为识别[21]、图像分类[22]、情感分析[23]等。基于关节点的行为识别研究中，Song 等[24]提出一种时空注意机制长短时期记忆网络(spatial-temporal attention long short-term memory, STA-LSTM)，用来学习堆叠 LSTM 层之间的注意权值。Zhang 等[25]提出了一种简单而有效的元素注意门(element-wise-attention gate, EleAttG)，可以很容易地添加到 RNN 块中(例如 RNN 层中的所有 RNN 神经元)，使 RNN 神经元具有注意能力。实验证明加入注意机制后的网络在识别效果上取得了更好的变现。

行为识别中并不是所有的帧都具有相同的重要性。有些帧捕获的信息意义不大，甚至带有与其他类型的行为相关的误导性信息，而有些帧则带有更多的辨别性信

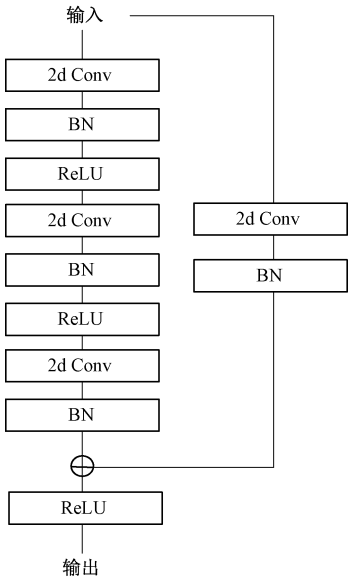


图 4　ResNet 中的残差块

Fig. 4　Residual block in ResNet

息。比如对“拥抱”与“靠近”这两类行为而言,“拥抱”发生之前两人会彼此“靠近”,而这些帧对于“拥抱”行为本身重要性并不大,且与“靠近”在前期有着相似性。因此本文算法在时序建模过程中加入全局注意机制,以赋予关键帧更多的权重,排除与识别的行为相关性较低的噪声帧的影响。

2) LSTM

LSTM 在学习输入特征序列之间的时间关系已被证明是非常有效的,并被广泛地用于具有时序信息的数据当中。

如图 5 所示,每个 LSTM 单元都包含一个单元状态,该状态保存前一单元中包含的信息,从而使得网络可以学习时间关系。这个单元状态是 LSTM 存储单元的一部分,其中遗忘门 f_t 和输入门 i_t 共同控制来自先前单元和隐藏状态的信息,用于生成新的信息并保存在新的当前状态 C_t 中。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

式中:σ 为 sigmoid 函数。这个新的单元状态连同一个输出门 O_t,用于生成单元的隐藏状态 h_t,如下所示:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = O_t * \tanh(C_t) \tag{6}$$

然后将单元状态和隐藏状态传递给连续的 LSTM 单元,允许网络学习长期依赖关系。

BLSTM 网络则是由两个朝相反方向前进的 LSTM 组成,从而最大程度地利用了来自过去和未来的关系中的可用上下文信息。这两个网络的输出相乘作为一个输出层。网络的输出同时考虑了前后的因素所得到,因此其全局化优势更加明显,更加具有鲁棒性。对于一个视频

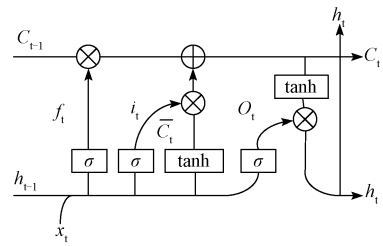


图 5　LSTM 单元模块

Fig. 5　LSTM modular unit

序列而言,如“靠近”和“拥抱”,如果按照 LSTM 来看,它是从前往后学习,视频的前一部分信息是这两个人在同时靠近,这两类行为就很相似,则难以快速学到后期的具有差别的信息。而 BLSTM 可以实现倒序进行信息的传递,那么当 BLSTM 读取完靠近的信息时,它的反向序列已经捕捉到了后半部分的内容信息(即拥抱相较于单纯的靠近还有两个人手部的信息),因此能够做出更准确的判断。本文选择两个连续的 BLSTM,每个 BLSTM 具有 512 个隐藏单元。另外,在网络中加入 Dropout,以避免训练时出现过拟合。

3) A-BLSTM

本文设计了 A-BLSTM 用于双人交互行为的时序建模,如图 6 所示,即在 BLSTM 的模型上加入 Attention 层。BLSTM 网络中会用最后一个时序的输出向量作为特征向量,然后送入 Softmax 中进行分类。加入 Attention 后,则是先计算每个时序的权重,然后将所有时序的向量进行加权和作为最终的特征向量,然后送入 Softmax 中进行分类。

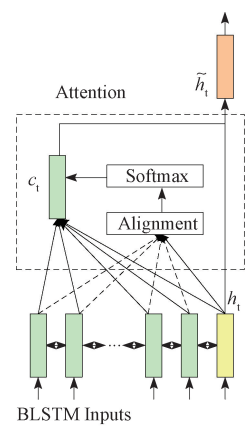


图 6　加入全局注意的 BLSTM

Fig. 6　BLSTM with global attention

加入全局注意机制的 BLSTM 的第 2 层的最终输出值 h_t,以及通过注意机制得到的上下文向量 c_t 用于生成最终的特征向量。如下所示:

$$\tilde{h}_t = \tanh(w_c[c_t; h_t]) \tag{7}$$

全局注意机制上下文向量是编码器所有隐藏状态输

出的加权和。其在 h_t 和每个隐藏状态值之间生成的对齐向量通过 softmax 层传递，然后用来表示上下文向量的权重。上下文向量的计算如式(8)所示：

$$c_t = \sum_{i=1}^{t} \bar{h}_i \left(\frac{\exp(h_t^{\top} \cdot \bar{h}_i)}{\sum_{i'=1}^{t} \exp(h_t^{\top} \cdot \bar{h}_{i'})} \right) \tag{8}$$

式中：$\bar{h}_i$ 表示第 i 个 BLSTM 隐藏状态的输出。

因此，采用的注意机制模块决定了每个帧的重要程度，将不同的注意力权重分配给不同的帧，更好地利用它们各自的辨别能力，更多地关注重要的帧。

3　实验结果与分析

3.1　数据集

NTU RGB+D 数据集[26]是目前最大的行为识别数据集。它是由多个 Kinect v2 相机捕获而成，包含 RGB，深度和骨架数据 3 种数据源。该数据集拥有超过 5 万个视频，包含共 60 类由 40 位年龄介于 10~35 岁的受试者完成的行为。包括 40 项日常行为、9 项医疗健康相关行为和 11 项双人交互行为。本文则采用骨架数据源中的 11 项交互行为，即击打、踢、推、拍背、指、拥抱、递物品、摸口袋、握手、靠近和远离，对本文所提算法进行评估，如图 7 所示。该数据集的官方评估方法协议有两种类型，交叉受试者(cross-subject，CS)和交叉视角(cross-view，CV)。CV 评估选择相机视角 2 和相机视角 3 的所有样本作为训练集，选择相机视角 1 的所有样本作为测试集。为应对实际场景中视角不同且多变的情形，本文采用 CV 协议进行评估。

本文实验是在 Ubuntu16.04 操作系统下进行，采用基于 python3.6 的深度学习框架 TensorFlow－1.8.0＋

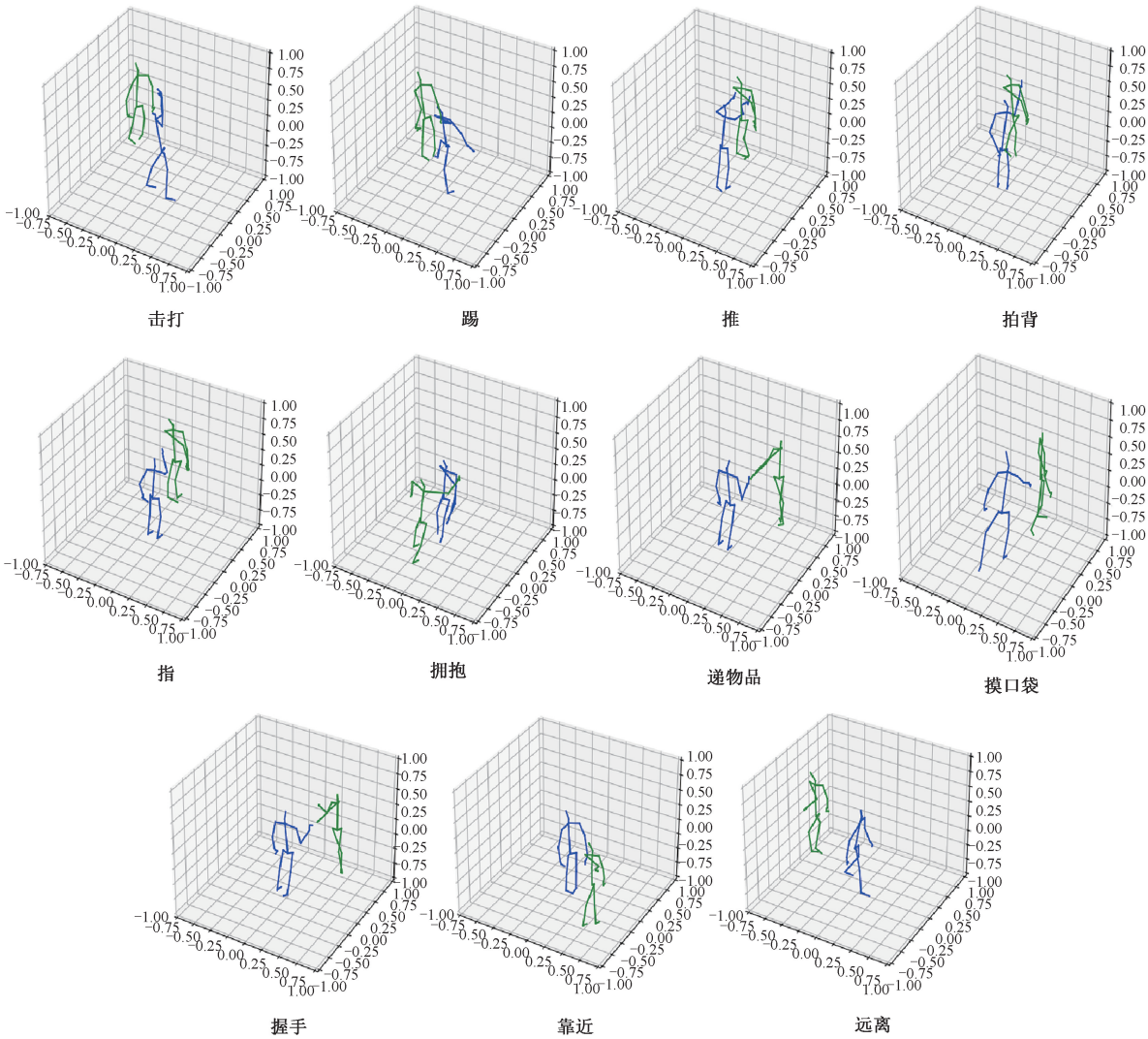


图 7　NTU RGB+D 骨架数据示例

Fig. 7　NTU RGB+D skeleton data examples

Keras2. 1. 6,GPU 为 NIVIDIA 1080 Ti 的深度学习环境。选择 Relu 作为激活函数,Adam 作为优化方法进行实验取优,其余参数均保持默认值,其中损失函数采用交叉熵损失函数,进行实验的训练与测试。

3.2　实验测试结果

本文将 NTU 数据集交互部分的关节点数据以行为类别进行划分,按照 CV 标准进行训练测试。本实验采用 50 次迭代进行训练,得到的训练集和测试集对应的准确率和损失函数的变化曲线如图 8 所示。

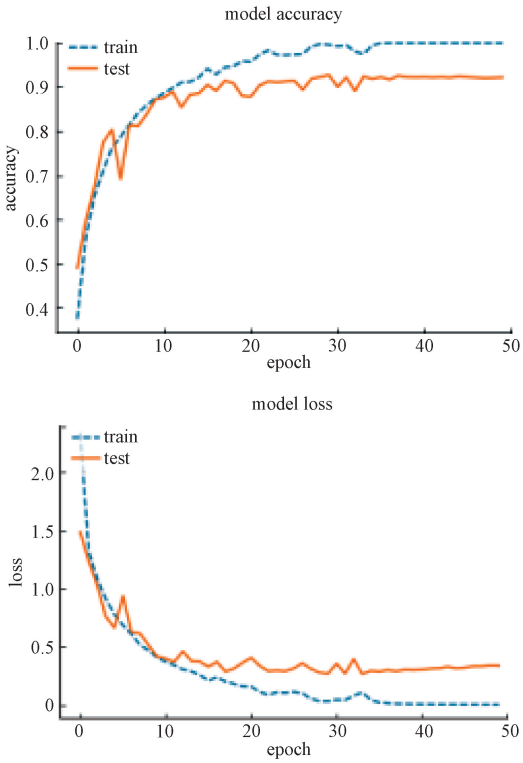


图 8　模型训练准确率和损失函数值

Fig. 8　Model training accuracy and loss value

由图 8 可知,模型最终收敛趋于稳定,得到的最高识别率为 90%。为了进一步分析模型的性能,生成图 9 所示的混淆矩阵。

由混淆矩阵可以看出,“递物品”和“握手”这两类行为容易混淆,主要因为在关节点层面,不能捕获到物品的信息,因此这两类行为存在一定的相似性。“摸口袋”和“拍背”也存在混淆,主要因为他们之间的交互只有其中一人执行动作,且运动变化均集中于手部,因而导致识别准确率相对较低。

3.3　对比实验

为了验证本文所提出算法,与在 NTU RGB+D 数据库中的双人行为数据上进行算法测试的文献所得到识别结果进行了比较和分析,结果如表 1 所示。

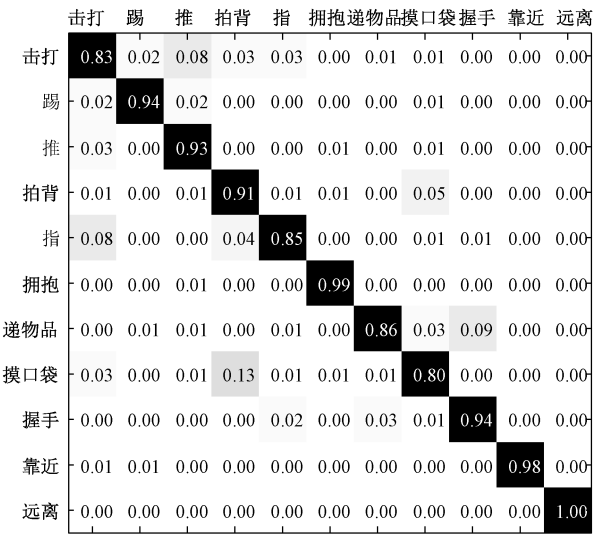


图 9　混淆矩阵

Fig. 9　Confusion matrix

表 1　在 NTU RGB+D 数据集上与其他方法准确率的对比

Table 1　The accuracy compared with other methods on the NTU RGB+D dataset

方法	准确率/%
Lie Group[11]	52. 8
Trust Gate LSTM[15]	77. 7
STA-LSTM[24]	81. 2
TS-LSTM[16]	82. 6
Ind-RNN[17]	88. 0
EleAtt-GRU[25]	88. 4
Clips + CNN + MTLN[13]	84. 8
TSRJI + CNN[14]	80. 3
Our method	90. 0

由表 1 可以看出,本文所提出的识别算法远优于传统方法[11]。相较于文献[13-14]只采用 CNN 的方法,所提算法的识别率有了很大的提升,证明了加入时序建模网络对视频的重要性和有效性。相较于单独采用 RNN 或 LSTM[15-17]的深度学习方法也得到了更好的识别效果,证明了对于具有时序建模能力的网络而言,输入的特征判别性越好,识别效果就会得到进一步地提升。相较于同样在网络中加入注意机制的方法[24-25],本文也获得了更高的识别率。主要原因在于首先从底层特征出发,充分地考虑了双人姿态信息和双人的交互信息。其次 CNN 与 A-BLSTM 的结合不仅提取到了更优的空间特征,还能有效地对关键帧进行时序建模,充分利用了两种网络各自的长处,从而达到优势互补的效果。最后基于简单的底层特征,在没有增加计算复杂度的基础上得到了理想的结果,具有在实际场景中应用的广阔前景,显示了本文算法的优越性。

4 结 论

本文提出了一种 CNN A-BLSTM 网络的双人交互行为识别模型,并在底层图像化过程中提出一种新的交互矩阵图表示。该算法框架的优势是在底层特征图像化过程中充分考虑个体关节点之间的依赖性和双人交互关系,在 CNN 提取到更好的空间特征的基础上还充分考虑了视角运动的时序关系,充分利用了这两种网络的优点,并且加入注意机制以赋予关键帧更多权重,排除与识别行为相关性较低的噪声帧影响。在实验部分,与现有算法进行比较分析,验证了所提方法的有效性。由于人体的运动是三维的,在将其进行二维图像化过程中难免会丢失一些信息,从而造成一定的局限性;只采用单一关节点数据源,对于本身相似性很大的行为难以进一步区分。考虑到人体骨架的拓扑结构本质上是一种基于图的结构,近期的研究表明图卷积网络在单人行为识别任务上优势明显,在未来的工作中,将研究构造双人交互的三维空间表示,通过图卷积网络提取更加完善的空间特征。另一方面,研究采用多源数据的有效融合方法以进一步区分相似行为,提高模型的鲁棒性和泛化能力。

参考文献

[1] WANG P, LI W, OGUNBONA P, et al. RGB-D-based human motion recognition with deep learning: A survey[J]. Computer Vision and Image Understanding, 2017, 171:118-139.

[2] 胡建芳,王熊辉,郑伟诗,等. RGB-D 行为识别研究进展及展望[J]. 自动化学报, 2019, 45(5): 829-840.
HU J F, WANG X H, ZHENG W SH, et al. RGB-D action recognition: Recent advances and future perspectives [J]. Acta Automatica Sinica, 2019, 45(5): 829-840.

[3] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]. IEEE International Conference on Computer Vision, 2015: 4489-4497.

[4] CARREIRA J, ZISSERMAN A, VADIS Q. Action Recognition? A new model and the kinetics dataset[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[5] DU Y, FU Y, WANG L. Skeleton based action recognition with convolutional neural network[C]. 3rd Asian Conference on Pattern Recognition, 2015: 579-583.

[6] WANG P, LI Z, HOU Y, et al. Action recognition based on joint trajectory maps using convolutional neural networks[C]. 24th ACM international conference on Multimedia, 2016: 102-106.

[7] 陈姝琪,曹江涛,赵挺, 等. 基于关节点数据的双人交互行为识别[J]. 电子测量与仪器学报,2020,34(6): 124-130.
CHEN SH Q, CHAO J T, ZHAO T, et al. Two-person interaction behavior recognition based on joint data[J]. Journal of Electronic Measurement and Instrument, 2020,34(6):124-130.

[8] HAN F, REILY B, HOFF W, et al. Space-time representation of people based on 3D skeletal data: A review[J]. Computer Vision and Image Understanding, 2017, 158: 85-105.

[9] LIU B, CAI H, JU Z, et al. RGB-D sensing based human action and interaction analysis: A survey[J]. Pattern Recognition, 2019, 94: 1-12.

[10] YUN K, HONORIO J, CHATTOPADHYAY D, et al. Two-person interaction detection using body-pose features and multiple instance learning[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012: 28-35.

[11] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3d skeletons as points in a lie group[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 588-595.

[12] JI Y, YE G, CHENG H. Interactive body part contrast mining for human interaction recognition[C]. IEEE International Conference on Multimedia and Expo Workshops, 2014: 1-6.

[13] KE Q, BENNAMOUN M, AN S, et al. A new representation of skeleton sequences for 3d action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3288-3297.

[14] CAETANO C, BREMOND F, SCHWARTZ W R. Skeleton image representation for 3d action recognition based on tree structure and reference joints[C]. 32nd SIBGRAPI Conference on Graphics, Patterns and Images, 2019: 16-23.

[15] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal lstm with trust gates for 3d human action recognition[C]. European Conference on Computer Vision, 2016: 816-833.

[16] LEE I, KIM D, KANG S, et al. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks[C]. IEEE International Conference on Computer Vision, 2017: 1012-1020.

[17] LI S, LI W, COOK C, et al. Independently recurrent neural network (indrnn): Building a longer and deeper rnn[C]. IEEE Conference on Computer Vision and

Pattern Recognition, 2018: 5457-5466.
[18] WANG P, YUAN C, HU W, et al. Graph based skeleton motion representation and similarity measurement for action recognition [C]. European Conference on Computer Vision, 2016: 370-385.
[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[20] JONES S, SHAO L. A multigraph representation for improved unsupervised/semi-supervised learning of human actions [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 820-826.
[21] DU W, WANG Y, QIAO Y. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos[C]. IEEE International Conference on Computer Vision, 2017: 3725-3734.
[22] ZHENG H, FU J, MEI T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition [C]. IEEE International Conference on Computer Vision, 2017: 5209-5217.
[23] YOU Q, JIN H, LUO J. Visual sentiment analysis by attending on local image regions[C]. AAAI Conference on Artificial Intelligence, 2017: 231-237.
[24] SONG S, LAN C, XING J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]. AAAI Conference on Artificial Intelligence, 2017: 4263-4270.
[25] ZHANG P, XUE J, LAN C, et al. Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks[J]. IEEE Transactions on Image Processing, 2019, 29: 1061-1073.
[26] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010-1019.

作者简介

赵挺,2018 年于辽宁石油化工大学获得学士学位,现为辽宁石油化工大学硕士研究生,主要研究方向为图像处理和模式识别。
E-mail:2442328581@ qq. com

Zhao Ting received his B. Sc. degree in 2018 from Liaoning Petrochemical University. Now he is a M. Sc. candidate at Liaoning Petrochemical University. His main research interests include image processing and pattern recognition.

曹江涛,2009 年于英国普茨茅斯大学获得博士学位,现为辽宁石油化工大学教授、硕士生导师,主要研究领域为智能方法及其应用、视频分析与处理等。
E-mail:cigroup@ 126. com

Cao Jiangtao received Ph. D. degree from University of Portsmouth in 2009. Now he is a professor and M. Sc. supervisor at Liaoning Petrochemical University. His main research interests include intelligent methods and their applications, video analysis and processing, etc.

姬晓飞,2010 年于英国普茨茅斯大学获得博士学位,现为沈阳航空航天大学副教授、硕士生导师,主要研究方向为视频分析与处理、模式识别理论等。
E-mail:jixiaofei7804@ 126. com

Ji Xiaofei received Ph. D. degree from University of Portsmouth in 2010. Now she is an associate professor and M. Sc. supervisor at Shenyang Aerospace University. Her main research interests include video analysis and pattern recognition theory, etc.