文章编号：2095-2163(2023)12-0129-04 中图分类号：TP391.4 文献标志码：A

基于骨架坐标的 LC 融合动作识别算法

冯 杰，郑建立
（上海理工大学 健康科学与工程学院，上海 200093）

摘 要：动作识别是计算机理解人类行为的关键技术，为了提高动作识别算法的时空特征提取能力，本文提出了一种融合 LSTM 和 CNN 的动作识别算法。该算法利用 LSTM 子网捕捉时间信息，采用 CNN 子网捕捉空间特征，然后融合特征进行动作识别。本文方法在 NTU RGB-D 数据集上，CS 验证的准确率达到 87.0%，CV 验证的准确率达到 91.5%。此外，针对动作时间长度不统一问题，同时对比了近邻插补和零向量插补方法，得到前者表现较优的结论。

关键词：动作识别；CNN；LSTM；人体骨架坐标

LC fusion action recognition algorithm based on skeleton coordinates

FENG Jie, ZHENG Jianli
(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Action recognition is a key technology for computers to understand human behavior. To improve the ability of extracting spatio-temporal features of action recognition algorithms, this paper proposes an action recognition algorithm that integrates LSTM and CNN. LSTM subnet captures temporal information, CNN subnet captures spatial features, and then fuses features for action recognition. The accuracy of CS verification on NTU RGB-D dataset and CV verification is 87.0% and 91.5% respectively. Aiming at the problem of non-uniform action time length, this paper compares the nearest neighbor interpolation and zero vector interpolation, and concludes that the former performs better.

Key words: action recognition; CNN; LSTM; Human skeleton coordinates

0 引 言

目前，人工智能领域迅速发展，许多研究人员都在关注如何让计算机理解人类行为，其中动作识别算法是关键性技术。动作识别正在越来越多地用于人们的日常生活，如：在人机交互[1]、VR 游戏领域，以及对人体动作的捕捉[2]；在安防领域，对人体行为的分析[3]，如：智能监控、肢体对抗等；在运动和康复领域，用于指导人的训练[4]。动作识别算法的输入数据通常为 RGB 视频、3D 骨架坐标数据等。RGB 视频作为输入，可直接实现端到端的动作识别，但 RGB 视频本身占用空间大，随着动作序列长度的增加，所需的计算复杂度进一步增加。同时在算法计算过程中，需要考虑背景等无关信息的干扰，进一步增加了识别难度。而基于骨架坐标数据进行动作识别，只需要考虑人体关节点的空间坐标关系，极大地减少了计算的数据量和复杂度。骨架数据只包含坐标，排除了背景等无关信息的干扰，对提高动作识别算法的准确率有所帮助。

本文提出一种基于骨架坐标的动作识别算法，算法包含两个并行子网，分别为 LSTM 子网（L 流）和一维 CNN（C 流）。其中 L 流捕捉动作时间特征，C 流捕捉关节点空间特征，最后融合两个子网的信息，输出强有力的动作识别结果。算法在公开数据集 RGB NTU-D 上达到领先水平。

1 LC 融合动作识别网络

1.1 算法框架

对于动作识别任务，不仅要考虑坐标之间的时间关系（运动顺序），还需要考虑空间关系（运动轨迹）。对于坐标之间的时间关系，LSTM 具有强大的捕捉时序特征的网络结构；对于坐标间的空间关系，本文采用一维卷积进行捕捉特征。网络的整体结构如图 1 所示。

基金项目：国家重点研发计划子课题（2020YFC2005802）。

作者简介：冯 杰（1994-），男，硕士研究生，主要研究方向：医学信息集成。
通讯作者：郑建立（1965-），男，博士，副教授，主要研究方向：医学信息集成。Email：zhengjianli163@163.com

收稿日期：2022-12-02

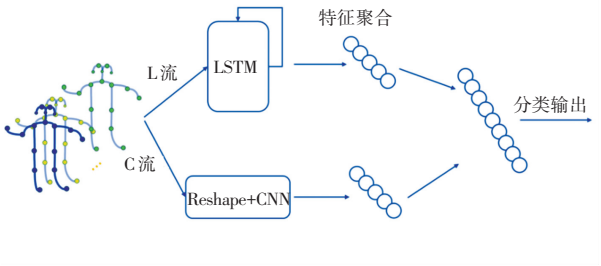

图 1 LC 融合动作识别网络

Fig. 1 LC fusion network

LSTM 子网中,本文采用 3 层 LSTM 连接进行特征提取,对于输入序列长度为 T 的样本,LSTM 结构可控制在每个时间步都进行输出,即输出为 $Y \in \mathbb{R}^{T\times d}$,或只在最后一个时间步进行输出 $Y \in \mathbb{R}^{1\times d}$。对于前两层 LSTM,选择每个时间步都输出一个向量,作为下一层 LSTM 的输入;对于最后一层 LSTM,只在最后一个时间步输出一个特征向量。具体结构如图 2 所示。

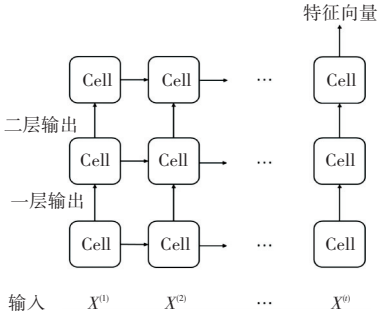

图 2 LSTM 子网络

Fig. 2 LSTM subnet

CNN 子网中,输入数据需要先经过维度重塑。由于一维卷积结构是作用在第一维度,对于数据集 $X \in \mathbb{R}^{T\times d}$ 来说,数据的第一个维度是时间维度,第二个维度是坐标维度即空间维度。C 流的目的是提取空间特征,因此需要将数据集的维度重塑为 $X \in \mathbb{R}^{d\times T}$。重塑之后,使用一维卷积重复作用在所有时序上,提取整个时间序列的空间信息,得到输出 $O \in \mathbb{R}^{d'\times T}$。此时,输出仍然是二维矩阵,其中包含每个时间步的信息;d' 为卷积处理后的每个时间步的特征向量长度,如式(1)所示:

$$d' = \frac{d + 2padding - f}{stride} + 1 \tag{1}$$

其中,$padding$ 为样本边界填充大小(本文选择不填充);f 为卷积核大小;$stride$ 为卷积步长。

然后,网络将融合矩阵中所有时刻的信息,即接入全连接层,最后得到 C 流的特征向量输出 $Z \in \mathbb{R}^{l\times 1}$,其中 l 是经过卷积提取后的特征向量的长度。具体架构如图 3 所示。

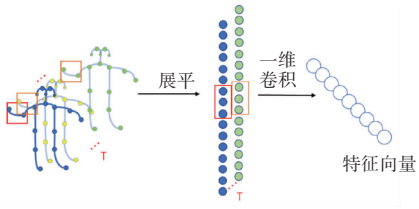

图 3 CNN 子网络

Fig. 3 CNN subnet

L 流和 C 流最终输出的特征向量可用于进行动作分类。将两个特征向量拼接在一起后,经过全连接层并由 softmax 激活函数,可得到类别的置信度输出。本文使用交叉熵损失函数指导模型参数学习,如式(2)所示。

$$Loss = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{2}$$

1.2 骨架动作时序填补

在 LSTM 结构中,对于输入序列要求每个样本都含有相同的时间步长度,因此对于原始数据样本中不同长度的动作序列,需要填补到相同的长度。近邻插值法填补过程如下:

假设原始动作序列为 $X = [X_1, X_2, \cdots, X_m]$,$X_i \in \mathbb{R}^{t_i\times 3d}$,$i \in [1,m]$,数据集总样本量为 m。其中 t_i 为样本 i 序列长度,d 为每个时间步的向量长度(在骨架数据中即关节点坐标数量),$3d$ 为每个关节点的 x、y、z 3 个坐标点。在原始数据中,每个样本的 t 不一定相等。设总样本集中序列最大长度为 T,则填补后的样本为 $X'_i \in \mathbb{R}^{T\times 3d}$,$i \in [1,m]$,每个样本拥有统一序列长度 T。近邻插值填补的过程如式(3)所示:

$$X'_i{}^{[k]} = X_i{}^{[floor(\frac{k}{T}t_i)]},\ k \in [1,T] \tag{3}$$

其中 $X_i{}^{[k]}$ 为原始第 i 个样本的第 k 个时间步,$X'_i{}^{[k]}$ 为填补后样本 i 的第 k 个时间步,$floor(x)$ 为向下取整函数,见式(4):

$$floor(x) = \max\{n \in Z \mid n \leqslant x\} \tag{4}$$

经过近邻插值填补后的样本,长度一致,可以送入后续动作识别模型中进行统一识别处理。

除了近邻插值填补外,还有一种比较简单的零向量填补。对于序列长度不到标准长度的样本,在序列前或序列后补零向量,使其达到标准长度。这种填补方式不需要计算插值,同时不会改变动作的快慢,适合需要考虑动作速度的任务。设原始样本为 $x_i = \{x_i^{[1]}, x_i^{[2]}, \cdots, x_i^{[t]}\}$,$x \in \mathbb{R}^{t\times d}$,$t$ 为未填补前的样本序列长度。填补后的样本为 $x_i = \{\mathbf{0}, \mathbf{0}, \cdots, x_i^{[1]}, x_i^{[2]}, \cdots, x_i^{[t]}\}$,$x_i \in \mathbb{R}^{T\times d}$,后填补之后的样本为 $x_i = \{x_i^{[1]}, x_i^{[2]}, \cdots, x_i^{[t]}, \mathbf{0}, \cdots, \mathbf{0}\}$,$x_i \in \mathbb{R}^{T\times d}$,$T$ 为统一序列长度。

使用零向量填补方式的样本,填补空间的值全

为 0 向量，这部分不携带任何信息，在模型训练时需要排除填补空间的干扰，即不计算填补空间部分的梯度。同时在模型正向推断时，填补部分应该被跳过，不参与计算。因此需要使用掩模（Mask）处理填补空间。掩模示意如图 4 所示。

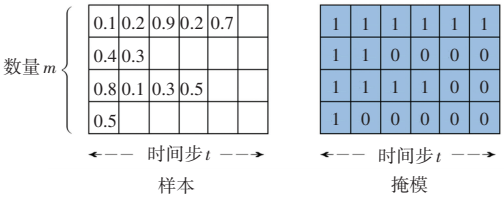

图 4　样本及其掩模示意图

Fig. 4　Sample and mask diagram

零向量填补实现简单，不需要计算，但对于序列较短的样本会产生大量的无效时间步；插值填补将动作弥散到最大时间长度上，对个体执行动作的快慢不敏感。

2　实验分析

实验采用 NTU RGB-D[5-6] 数据集对本文提出的 LC 融合动作识别算法进行验证。该公开数据集使用微软 Kinect2.0 设备采集，每个样本包含 RGB 视频、深度图序列、3D 骨架坐标数据和红外视频，本文只使用 3D 骨架坐标数据。图 5 为 3D 骨架坐标的可视化图中，“站起”动作里的中间三帧。

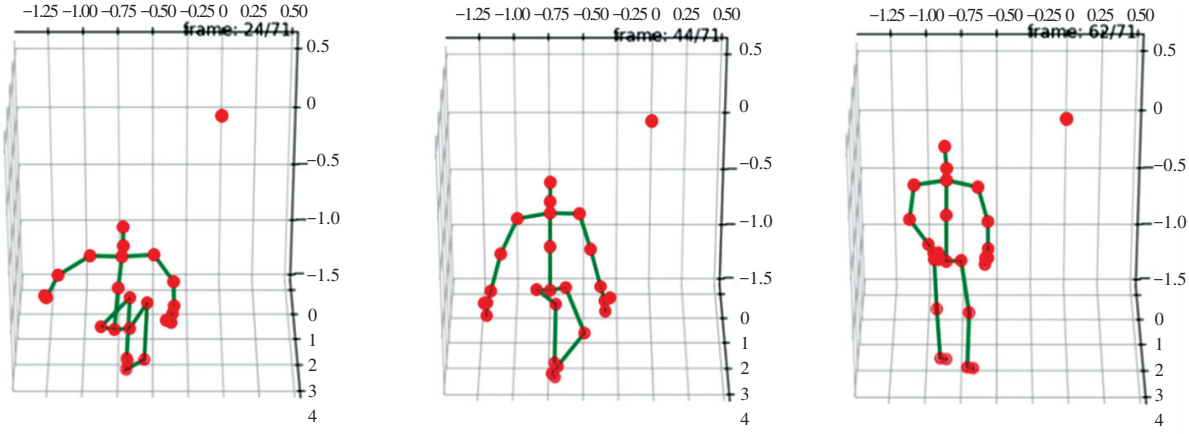

图 5　站起动作骨架可视化

Fig. 5　Visualization of standing up action

NTU RGB-D 数据集的骨架点共有 25 个，包含 40 个参与数据采集的个体，每个参与个体会进行 60 种不同的动作，动作会表演多次，同时有 3 个不同的相机角度进行拍摄，其中 3 个摄像头的垂直高度是一致的，水平角度分别是 -45°、0° 和 45°。最后，设置不同的相机高度和距离以增加视角的多样性，组成了共 56 880 个动作样本。本次实验与 NTU 动作识别竞赛规则相同，在划分训练集和测试集上有两种标准：CS（Cross-Subject）为交叉个体验证，指样本的 40 个个体中，编号为[1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38]的个体作为训练集，剩余的作为测试集，训练集共 40 320个样本，测试集共16 560个样本；CV（Cross-View）为交叉视野验证，指按摄像头来划分训练集和测试集，摄像头 1 采集的样本作为测试集，摄像头 2 和 3 作为训练集，样本数分别为 18 960 和37 920。数据集的示例如图 6 所示，文件名即为动作标签。其中，S 表示设置号（1-17）、C 表示相机 ID（1-3）、P 表示参与者 ID（1-40）、R 表示动作执行次数（1-2）、A 表示动作类别（1-60）。

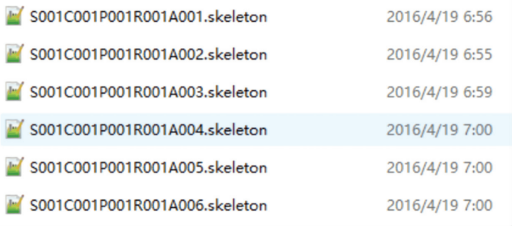

图 6　NTU RGB-D 骨架数据集示例

Fig. 6　Example of NTU RGB-D skeleton dataset

本次实验样本插补序列长度 T 取数据集中最大视频长度 300 帧，采用 $\alpha = 0.001$ 作为初始学习率，在迭代训练 200 个 epoch 之后，得到结果见表 1。

表 1 中 LC Fusion 网络为本文所提出的算法，其中 Zeros Pad 代表使用零向量填补，NN Pad 代表使用最近邻插值进行填补。从表中可以看出，LC 融合动作识别算法能够有效地同时捕捉关节的时间信息和空间信息，与其它算法相比效果显著，在 CS 验证的准确度超过了 85%，在 CV 验证的准确度超过了 90%。同时可以看出，最近邻插值填补法的效果优于零向量填补。直观来讲，动作识别任务中，个人的动作快慢并非最核心的因素，个体的整体运动趋势才是决定动作的关键要素。最近邻插补将整个动作

弥散到最大时间尺度上,能有效利用网络提取特征的能力,而零向量插补在动作序列较短的样本上,使网络跳过了许多时间步的计算,没有充分利用所有的时间步,因此效果相比前者较差。但与其它算法对比,零向量插补的方法准确率也有所领先,证明了本文 LC 融合识别算法的有效性。图 7 给出了算法预测的 60 个类别在 CS 测试集上的 ROC 曲线,最小 AUC 为 0.96,最大 AUC 为 1.0,平均 AUC 为 0.99。

表 1 NTU RGB-D 实验准确率对比

Table 1 Comparison of accuracy of NTU RGB-D experiment

方法	CS/%	CV/%
HBRNN-L[7]	59.1	64.0
Dynamic Skeletons[8]	60.2	65.2
Part-aware LSTM[9]	62.9	70.3
ST-LSTM + Trust Gate[10]	69.2	77.7
STA-LSTM[11]	73.4	81.2
GCA-LSTM[12]	74.4	82.8
URNN-2L-T[13]	74.6	83.2
Clips+CNN+MTLN[14]	79.6	84.8
ESV (Synthesized + Pre-trained)[15]	80.0	87.2
IndRNN (6 layers)[16]	81.8	88.0
ST-GCN (with JPD)[17]	83.4	88.8
LC Fusion (Zeros Pad)	85.8	88.7
LC Fusion (NN Pad)	87.0	91.5

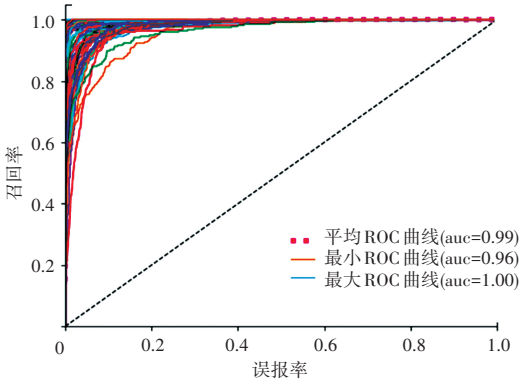

图 7 LC 融合算法 ROC 结果

Fig. 7 ROC curve of LC fusion algorithm

3 结束语

本文提出的 LC 融合动作识别算法,通过两个子网分别捕捉骨架动作数据的空间特征和时间特征。经过融合两个子网的特征后,使算法具有强大的时空特征提取和分类能力。对于骨架动作数据时长不统一问题,通过实验验证了近邻插补的效果优于零向量插补。本文算法在公开数据集 NTU RGB-D 上具有较优效果。

参考文献

[1] 唐彪,樊启润,孙开鑫,等. 人体姿态识别算法在视觉人机交互中的应用[J]. 计算机测量与控制,2019,27(7):242-247.

[2] 张继凯,顾兰君. 基于骨架信息的人体动作识别与实时交互技术[J]. 内蒙古科技大学学报,2020,39(3):266-272.

[3] 马子健,林雨衡,王志强,等. 封闭环境下人体姿态识别及打架行为监测[J]. 计算机应用,2021,41(S2):214-220.

[4] 闫航,陈刚,佟瑶,等. 基于姿态估计与 GRU 网络的人体康复动作识别[J]. 计算机工程,2021,47(1):12-20.

[5] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1010-1019.

[6] LIU J, SHAHROUDY A, PEREZ M, et al. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2684-2701.

[7] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1110-1118.

[8] HU J F, ZHENG W S, LAI J, et al. Jointly learning heterogeneous features for RGB-D activity recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5344-5352.

[9] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1010-1019.

[10] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal lst with trust gates for 3d human action recognition[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 816-833.

[11] SONG S, LAN C, XING J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2017:4263-4270.

[12] LIU J, WANG G, HU P, et al. Global context-aware attention lstm networks for 3d action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1647-1656.

[13] LI W, WEN L, CHANG M C, et al. Adaptive RNN tree for large-scale human action recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1444-1452.

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

[15] LIU M, LIU H, CHEN C. Enhanced skeleton visualization for view invariant human actionrecognition[J]. Pattern Recognition, 2017, 68: 346-362.

[16] LI S, LI W, COOK C, et al. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5457-5466.

[17] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 7444-7452.