Computer Science and Application 计算机科学与应用, 2024, 14(1), 113-122
Published Online January 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.141013

Hans汉斯

基于循环神经网络的人体动作在线识别方法

李高丰

洛阳师范学院，物理与电子信息学院，河南 洛阳

收稿日期：2023年12月25日；录用日期：2024年1月24日；发布日期：2024年1月31日

摘　要

由于人体运动的非刚性以及表观特征的多样性，现实场景中对人体动作的准确识别较难。本文提出一种基于循环神经网络的人体动作在线识别方法，把人体关节点作为主要特征，利用深度学习进行动作识别建模。采用循环神经网络学习连续动作的时序关联信息，引入随机权重共享的注意力机制提高训练准确率，避免过拟合现象，实现对人体动作的在线识别。通过UCF数据集进行训练和测试，本文方法达到较高的准确率和稳定性，表明了基于循环神经网络的动作识别模型对现实场景人体动作的在线识别是有效的。

关键词

动作识别，姿态检测，循环神经网络，注意力

Research of Human Action Online Recognition Based on Recurrent Neural Network

Gaofeng Li

College of Physics & Electronic Information, Luoyang Normal University, Luoyang Henan

Received: Dec. 25th, 2023; accepted: Jan. 24th, 2024; published: Jan. 31st, 2024

Abstract

Due to the non-rigidity of human motion and the diversity of apparent features, it is difficult to accurately identify human action in real scenes. In this paper, an online recognition method of human action based on recurrent neural network is proposed. The human joint points are used as the main features. The deep learning is used to model the action recognition. The recurrent neural

文章引用: 李高丰. 基于循环神经网络的人体动作在线识别方法[J]. 计算机科学与应用, 2024, 14(1): 113-122.
DOI: 10.12677/csa.2024.141013

network learns the temporal correlation of continuous actions and realizes online recognition of human actions, in which the attention mechanism is introduced to improve the training accurate rate and decrease over fitting. Through the training and testing of UCF dataset, the proposed method achieves high accuracy and stability, which shows that the action recognition model based on recurrent neural network is effective for online recognition of human motion in real scenes.

Keywords

Action Recognition, Pose Estimator, Recurrent Neural Network, Attention

1. 引言

人体动作识别是人体运动分析的基础，广泛应用在视频分析、人机交互、自动驾驶、公共安全等领域，有着重要的研究意义和应用价值[1]。随着大量视频数据的增长，尽管数据处理能力有较大提高，对人体动作识别的准确性、实时性也提出了更高的要求。由于人体非刚性的运动和视觉特征变化的多样性，以及现实场景复杂性的影响，对人体动作的智能识别仍然存在许多困难。人体动作既有整体运动，又有局部移动，每个动作都是由关节点联系在一起的时空运动。由于类内动作的差异和空间角度的变化，导致运动特征空间具有复杂多变的形式。三维动作识别正在成为研究的方向，基于深度图像的数据集也在增加，但是二维 RGB 视频依然是数据的主要来源，在现实应用中更加普遍。在二维视角下人体视觉特征会产生旋转、缩放、形变、遮挡等变化，对看似简单的类内动作识别带来影响[2] [3]。在现实场景中，人与环境的交互情况对人体动作识别会产生一定影响。

动作特征是典型的空间特征与时间特征的融合，常用的描述有特征轨迹、容积法、光流场等。Qin 等提出基于时空特征点轨迹的动作识别方法，采用 KLT 跟踪时空局部特征，把跟踪得到的轨迹作为特征向量，利用多核学习方法进行分类识别[4]。Mekruksavanich 借助可穿戴传感器，提出一种多模态训练方法，实现了对运动相关动作识别[5]。Sevilla-Lara 提出把光流集成于动作识别之中，提高了动作识别的准确度[6]。传统的识别方法提取复杂的特征向量，送给训练好的分类器进行识别。特征提取和动作识别是两个分离的过程，由于动作特征的时空性，增加了提取动作特征的复杂程度。对不同目标的同一动作，甚至同一目标的类内动作，都很难确定特征向量的一致性。训练和测试的数据集一般都是手工分割的特定序列，受特征向量和环境的影响，产生模型泛化能力较弱的问题。

深度学习利用多层神经网络来学习复杂的特征，在图像处理、目标识别中表现出了良好的性能。由于数据量的增加，以及计算能力的提高，基于卷积神经网络的方法在动作识别中的运用越来越多[7]。Simonyan 等人提出一种空间卷积结合时域光流的双流结构模型，取得较好的性能[8]。Cao 基于训练图提取 CNN 特征，结合图像 RGB 数据和注意力算法，利用 LSTM 训练分类[9]。ZHA S 直接对视频进行卷积特征提取，利用 SVN 进行动作分类[10]。深层网络 AlexNet、VGG 模型通过增大网络的深度来获得更好的训练效果[11] [12]。GoogLeNet 通过引入 Inception 加宽网络结构，使网络结构稀疏连接，降低了层数增加带来的负作用[13]。Huang 提出复杂的 DsenseNet，将所有网络层两两进行连接，每一层都接受它前面所有层的特征作为输入，使网络模型大量密集的连接[14]。Wei 提出改进的 GooLeNet，对 softmax 进行缩小和细化，并增加分类输出层，具有较好的泛化能力[15]。随着卷积网络深度和宽度的不断增加，目

标特征经过层层处理其分布会逐渐产生差异，容易引起梯度消失、过拟合、训练缓慢等问题。特别是在动作识别任务中，受人体非刚性体运动和视角变化的影响更为突出。

为了降低模型复杂程度，提高训练准确率，并以常见的二维视频为处理对象，本文提出了一种基于循环神经网络人体动作在线识别方法，以人体关节点为主要特征进行逐帧处理，利用多层循环神经网络模型进行序列的动作特征识别。在模型中引入注意力机制，对帧间动作特征进行加权处理，增加与识别动作相关特征的权值比重，提高动作识别的准确率，降低过拟合风险。

2. 模型

直接对图像或序列提取运动特征，需要较深的网络层次和大量单元。动作是一种时间序列信号，可以看成特定姿态的集合。在人对动作的理解中，姿态是密切相关。有的动作可用固定的姿态表征，有的动作需要多个姿态严格按时序关系表征。姿态在人体动作识别中起到了关键作用，而且不会产生大量的数据计算。本文的模型以人体关节点作为识别动作的姿态特征，以时间序列的形式送入循环神经网络(Recurrent Neural Network, RNN)进行模型参数学习。

通过对视频均匀采样得到帧画面，采用卷积网络 CNN 模块初步处理，然后提取姿态信息，把所需的关节点及表观特征送入多层循环神经网络，进行动作识别训练。人体姿态利用关节点信息，在加上肢体表观的特征组成，包括方向、梯度等信息。在多层循环神经网络中引入注意力机制，学习序列中人体动作的帧间运动特征，最后通过 Softmax 输出识别的动作类型。模型框架见图 1 所示。

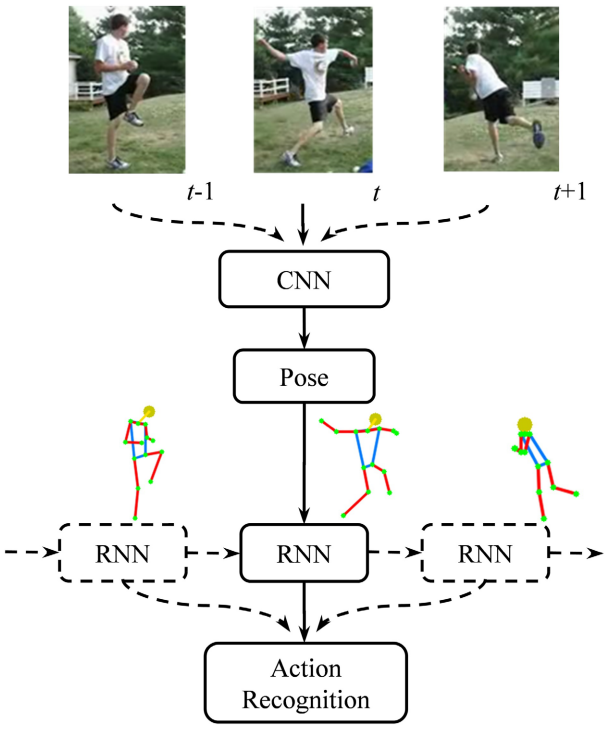

Figure 1. Model framework
图 1. 模型框架

3. 模型

通过 CNN 模块提取图像特征后，送入 Pose 模块进行提取姿态信息，简化人体关节模型，进行姿态估计。关节点和人体主要部件表观特征作为进入循环神经网络模块的数据。

3.1. CNN 模块

从帧图像中提取动作的空间特征，使用轻量化的卷积网络，采用深度可分离卷积方法，可以减少网络参数的数量，降低训练的计算量。输入网络的维度是 224 × 224 × 3，共有 12 层。第 1 层是标准的卷积层，有 24 个卷积核，步长是 2。从第 2 层开始，进行深度可分离卷积，最后一层的输出维度是 28 × 28 × 384。见表 1 所示，Conv dp 表示深度可分离卷积，包含 3 × 3 的深度卷积和 1 × 1 的逐点卷积。

Table 1. Architecture of CNN
表 1. CNN 网络结构

序号	卷积层	卷积核	输入尺寸	步长
L0	Conv	3 × 3 × 3 × 24	224 × 224 × 3	2
L1	Conv dp	3 × 3 × 2	112 × 112 × 24	1
L2	Conv dp	3 × 3 × 2	112 × 112 × 48	2
L3	Conv dp	3 × 3 × 1	56 × 56 × 96	1
L4	Conv dp	3 × 3 × 2	56 × 56 × 96	2
L5	Conv dp	3 × 3 × 1	28 × 28 × 192	1
L6	Conv dp	3 × 3 × 2	28 × 28 × 192	1
L7	Conv dp	3 × 3 × 1	28 × 28 × 384	1
L8	Conv dp	3 × 3 × 1	28 × 28 × 384	1
L9	Conv dp	3 × 3 × 1	28 × 28 × 384	1
L10	Conv dp	3 × 3 × 1	28 × 28 × 384	1
L11	Conv dp	3 × 3 × 1	28 × 28 × 384	1

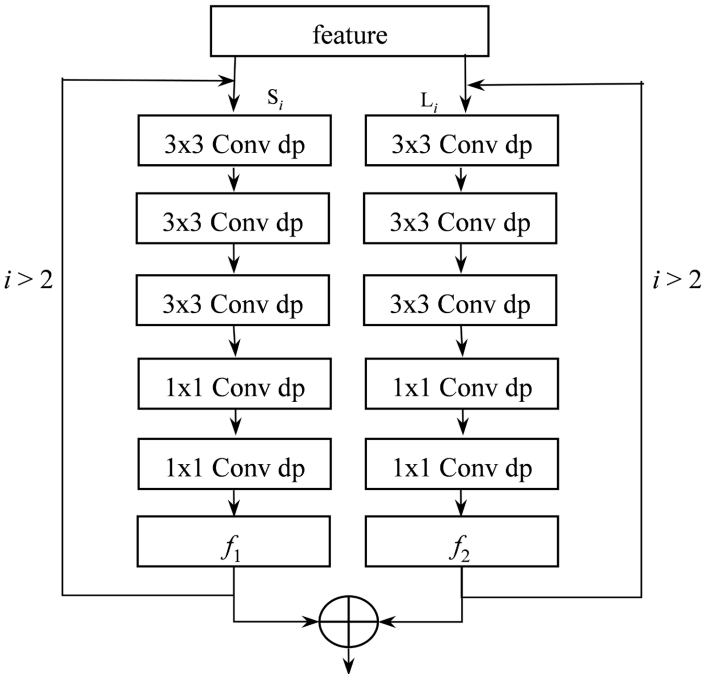

Figure 2. Pose extraction
图 2. 姿态提取

L1 层将输入数据按特征维度 24 分组，每一组进行 3 × 3 卷积，获得空间特征。在此基础上做 2 个 1 × 1 的逐点卷积，获得每个点的在各个卷积核下的特征。L2 层输出为维度是 112 × 112 × 48，得到 48 个相同分辨率的特征向量组。L3 层经过 2 × 2 池化后，再加上 L7 层和 L11 层组成 CNN 网络的输出特征向量。其它各层处理方法相同。此网络的计算量与相同容量的标准卷积网络比较，大约可以降低的 80%，虽然牺牲了一定的精度，但适合视频序列中运动特征的提取。

3.2. 姿态提取

CNN 模块提取特征后，通过 Pose 模块进行训练，每个阶段均计算关节响应和肢体表观特征，如方向、梯度。计算过程见图 2 所示，S 分支计算人体关节点的特征值，用于学习对应的置信度；L 分支计算人体部件向量，即关节点之间的关联度。循环 6 个阶段，每个阶段都深度可分离卷积代替标准卷积，而且 2~6 阶段的卷积尺寸是 3 × 3，降低计算量。

对人体关节模型进行简化，共有 14 个节点，重点分析姿态中的关键节点，如头、躯干、四肢等。f_1，f_2 分别是两个分支的损失函数，处理姿态的估计问题。总的代价函数由两个分支在每个阶段的损失之和组成，见公式(1)。

$$f = \sum_{t=1}^{N}\left(f_S^t + f_L^t\right) \tag{1}$$

其中，

$$f_S^t = \sum_{j=1}^{J}\sum_p W(p)\cdot\left\|S_j^t(p) - S_j^*(p)\right\|_2^2 \tag{2}$$

$$f_L^t = \sum_{c=1}^{C}\sum_p W(p)\cdot\left\|L_j^t(p) - L_j^*(p)\right\|_2^2 \tag{3}$$

公式(2)表示关节点的误差，公式(3)表示肢体特征的误差，用 W 用来对关节点加权，减小错误检测时对损失函数的影响。将所有阶段的损失函数相加，作为总损失函数。

4. 动作识别

提取人体姿态特征，可以组成时间序列向量。建立多层循环神经网络进行学习，在其中嵌入注意力网络层，提升样本训练的准确率，实现在线动作序列的识别。动作识别模块见图 3，序列数据首先通过批再归一化(BReN)处理，使样本分布统一，取得了较好收敛特性。多层循环神经网络处理归一化后的时间序列数据，激活函数采用 ReLU 函数。输出特征进入随机共享权重的注意力层(Attention Layer)，在训练过程中可以自动加强特征学习的权重，从而提高训练的效率。

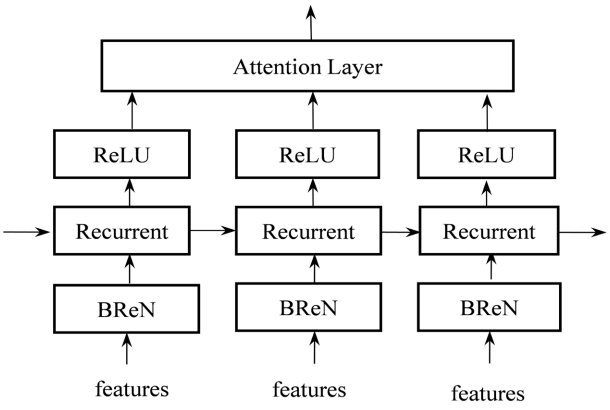

Figure 3. Action recognition
图 3. 动作识别

4.1. 循环神经网络

循环神经网络是一类用于处理时间序列数据的神经网络。常规的 RNN 网络处理较长序列存在梯度消失和长期依赖问题。动作与姿态关系密切，有的动作只用一帧就可以确定，比如走、跑、跳等。但复杂的动作序列持续较长，要通过上下文的联系才能识别。长短期记忆神经网络(Long Short-Term Memory, LSTM)是一种特殊的 RNN，通过加入门控制单元记忆长期信息，解决长序列在训练过程中梯度衰减过快的问题。

使用三层 LSTM 网络结构叠加，根据序列相邻间隔帧的姿态特征学习识别动作。由于训练时的样本一般是分割好的，现实场景的序列不一定是分割好的视频，多层结构利用上下文可以实现在线识别。LSTM 网络层结构见表 2，每一层均进行批再归一化操作 BReN 和 Dropout。用 BReN 进行样本的归一化处理，Dropout 随机舍弃部分神经元的权重，实现正则化处理作用。

Table 2. Recurrent neural networks
表 2. 循环神经网络

层	输出
Input	(Batch, Step, 34)
LSTM + BReN + Dropout	(Batch, Step, 128)
LSTM + BReN + Dropout	(Batch, Step, 128)
LSTM + BReN + Dropout	(Batch, Step, 128)
Attention +Flatten	(Batch, Step*128)
Dense + BReN + Dropout	(Batch, 128)
Dense + BReN + Dropout	(Batch, 64)
Softmax	(Batch, 32)

4.2. 注意力机制

在多层 LSTM 网络输出后加入注意力层，对重要特征进行加权，提高训练效率和准确率，降低复杂网络结构带来的过拟合风险。注意力机制在训练过程中，先计算特征的权值，然后根据样本输出加权求和，不断加大响应较大的特征权值，相应特征对识别的贡献也就增大。

复杂的网络结构和时序深度的增加，容易导致过拟合问题。注意力机制的产生可以有效缓解这个矛盾。在 LSTM 网络层之后加入注意力层，而不是直接对关节点特征引入注意力，目的是强化时间序列处理后的特征在动作识别中的作用。经过 LSTM 网络层得到的特征，经过注意力层的学习受到识别动作更多关注。

注意层的输出：

$$y_t = f_{att}\left(y_{t-1}, h_t, c_t\right) \tag{4}$$

其中，y_{t-1} 表示上一时刻的注意力输出，h_t 表示循环神经网络的隐状态，c_t 表示特征加权，f_{att} 表示注意力层函数。

在特征加权过程中，直接共享注意力权重可进一步降低大量参数的计算，但对于特征的表达能力减弱。本文给出一种随机共享权重的方法对特征加权：

$$c_t = \sum_{i=1}^{T} a_{t,i} h_i \tag{5}$$

其中，$a_{t,i}$ 表示注意力的权值。随机按一定概率给每个维度共享权重。公式(6)中 r 服从伯努利分布，$a_{t,i}$ 以概率 p 共享各维度权值的平均值，见公式(7)。

$$r_j^l \sim \text{Bernoulli}(P) \tag{6}$$

$$a_{t,i} = \begin{cases} a_{t,i}^k, & r_j^l = 0 \\ \dfrac{\sum_k a_{t,i}^k}{N}, & r_j^l = 1 \end{cases} \tag{7}$$

注意力层函数 f_{att} 的功能实现见图 4 所示。Scale 进行数据维度调整，Connect 是全连接神经单元，Mean 进行均值计算，Reduction 是在随机概率 p 下进行权值均值共享，最后对特征进行加权处理。权值随机共享的方法提升了显著性特征的权重，有利用于训练效率的提高，降低过拟合的出现。

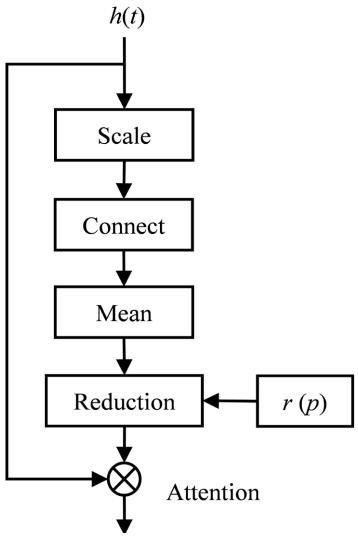

Figure 4. Function of attention layer
图 4. 注意力层功能

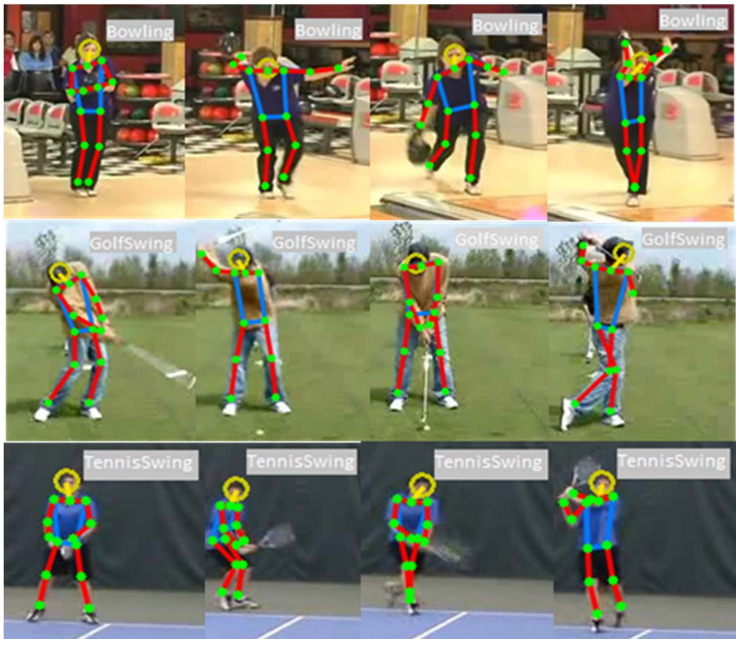

Figure 5. Result of action recognition
图 5. 动作识别结果

5. 实验

本文进行了模型的训练和测试。从 UCF 视频动作库中选择部分典型场景的动作视频作为样本，进行 32 类动作的识别。识别的实例见图 5 所示，图中有人体动作的关节点，以及对动作类型的分类标签。图 5 给出的 3 个识别动作，分别是 Bowling，GolfSwing，TennisSwing。每一行选取了 4 帧画面，可以准确检测到人体动作的关节点，正确识别了动作类型。

样本是 4000 个时间长短不同的视频，均匀采样 60 帧，分别组成训练集和测试集，进行模型的训练和交叉验证。样本训练和测试的准确率曲线，以及损失函数曲线见图 6。训练集的准确率最高可达 0.98，损失函数的下降有少量起伏，测试集的准确率也能同步上升到 0.97。本文提出的模型具有较好的泛化能力，减少了过拟合的现象。

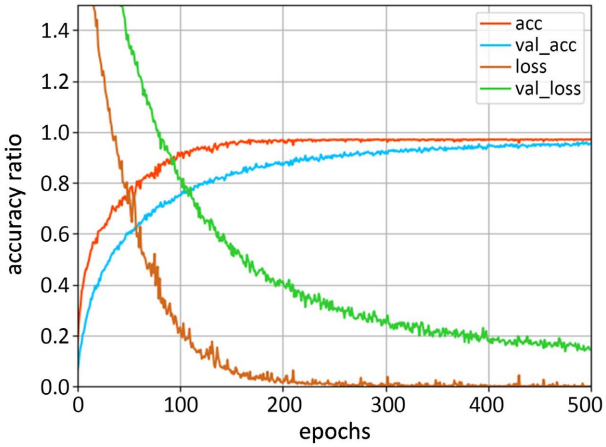

Figure 6. Accuracy and loss curves of training and validation
图 6. 样本训练与测试曲线

注意力机制嵌入在循环神经网络中，起到了显著的作用，图 7 给出了模型在有注意力层和无注意力层时训练集的准确率变化曲线对比。当没有注意力层时，准确率上升缓慢。加入注意力层后，准确率上升较快，提高了模型训练效率和准确率。经过循环神经网络后的特征值与动作识别密切相关，图 8 给出了 128 维特征向量在注意力层的权值比重分布，权值比重提升了对动作序列的表达能力。

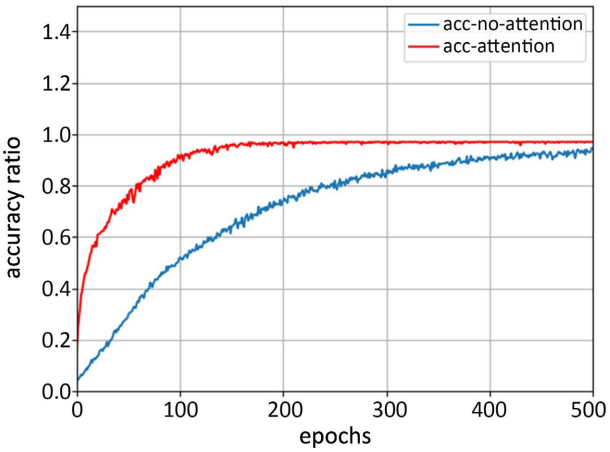

Figure 7. Comparison of attention layer
图 7. 注意力层对比

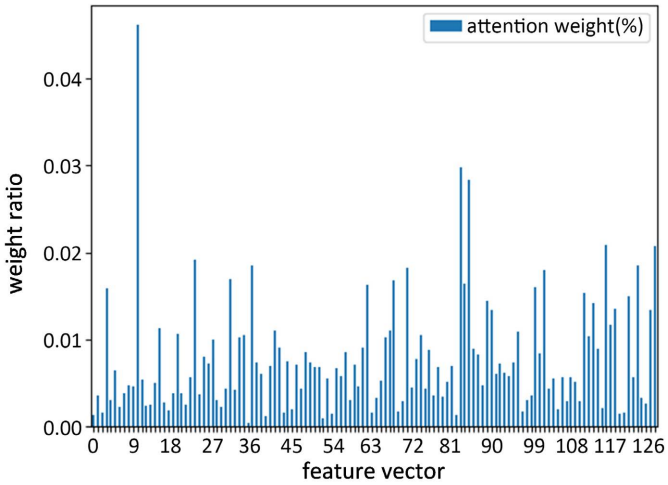

Figure 8. Weight ratio of attention layer
图 8. 注意力层权值比重

6. 结论

本文提出一种基于循环神经网络的人体动作在线识别方法，通过逐帧提取关节点特征，用循环神经网络处理时间序列数据，提出随机共享权重的注意力机制，实现了人体动作的在线识别。动作识别在实际中是困难的，不仅由于动作的复杂度不同，人体关节的非刚性等因素，还受到人与环境的交互影响，以及视频分辨率和采样率的差异。动作识别可以借助对环境的理解，人与其他事物的交互关系在动作识别中起到更直接的作用。由于每个视频的长短不同，采样的帧数会影响动作特征描述的信息量差异。从关键帧的姿态入手，结合辅助帧可以更有效的描述动作。在循环神经网络层后加入自定义的注意力层，通过训练自动对提取出的特征进行加权处理。提取特征的含义还不明确，但能在更高维度上对动作序列表征。今后，可以考虑加入对环境信息的学习与识别，把动作识别与背景建模、人与环境的交互结合起来，利用场景理解更好的实现动作在线识别。

参考文献

[1] 罗会兰, 童康, 孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报, 2019, 47(5):1162-1173.

[2] 邵志文, 周勇, 谭鑫. 基于视觉的人体动作识别综述基于深度学习的表情动作单元识别综述[J]. 电子学报, 2022, 50(8): 2003-2017.

[3] 包本刚. 融合多特征的目标检测与跟踪方法[J]. 电子测量与仪器学报, 2019, 33(9): 93-99.

[4] 秦磊, 胡琼, 黄庆明, 等. 基于特征点轨迹的动作识别[J]. 计算机学报, 2014, 37(6): 1281-1288.

[5] Mekruksavanich, S. and Jitpattanakul, A. (2022) Multimodal Wearable Sensing for Sport-Related Activity Recognition Using Deep Learning Networks. *Journal of Advances in Information Technology*, **13**, 132-138. https://doi.org/10.12720/jait.13.2.132-138

[6] Sevilla-Lara, L., Liao, Y.Y., Guney, F., *et al*. (2017) On the Integration of Optical Flow and Action Recognition. In: Brox, T., Bruhn, A. and Fritz, M., eds., *Pattern Recognition*. *GCPR* 2018. Springer, Cham.

[7] 朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848-857.

[8] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. https://arxiv.org/abs/1406.2199

[9] 曹晋其, 蒋兴浩, 孙锬锋. 基于训练图 CNN 特征的视频人体动作识别算法[J]. 计算机工程, 2017, 43(11): 234-238.

[10] Zha, S., Luisier, F. and Rews, W., *et al*. (2015) Exploiting Image-Trained CNN Architectures for Unconstrained Video Classification. https://arxiv.org/abs/1503.04144

[11] Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Conference and Workshop on Neural Information Processing Systems*, Lake Tahoe, Nevada, 3-8 December 2012.

[12] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556

[13] Szegedy, C., Liu, W., Jia, Y.Q., *et al*. (2014) Going Deeper with Convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 7-12 June 2015. https://doi.org/10.1109/CVPR.2015.7298594

[14] Huang, G., Liu, Z., Maaten, L.V.D., *et al*. (2017) Densely Connected Convolutional Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 21-26 July 2017. https://doi.org/10.1109/CVPR.2017.243

[15] Wei, L.R., Yue, J., Zhu, H., *et al*. (2019) Human Action Recognition Method Based on Deep Neural Network. *Journal of University of Jinan* (*Science and Technology*), **33**, 215-223.