自适应融合 RGB 和骨骼特征的行为识别

郭伏正，孔军*，蒋敏
江南大学模式识别与计算智能国际联合实验室，江苏 无锡 214122

摘要 传统的基于 RGB 和骨骼特征的行为识别算法，普遍存在两种特征互补性不足及视频关键时序性不强等问题。为解决这一问题，提出一种自适应融合 RGB 和骨骼特征的行为识别算法。首先，面向 RGB 图像和骨骼图像，联合双向长短时记忆（LSTM）网络和自注意力机制提取两者的时空特征；然后，构建自适应权重计算网络（AWCN），并以两者的空间特征为输入计算出自适应权重；最后，利用自适应权重得到上述时空特征的融合特征，实现了最终的动作分类。通过在 Penn Action、JHMDB 和 NTU RGB-D 人体行为数据集上与现有的方法进行比较，实验结果表明，本文算法有效地提高了行为识别精度。

关键词 机器视觉；行为识别；姿态估计；自适应权重计算网络；长短时记忆网络；自注意力
中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP57.201506

Action Recognition Based on Adaptive Fusion of RGB and Skeleton Features

Guo Fuzheng, Kong Jun*, Jiang Min
International Joint Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract In this paper, we proposed an action recognition algorithm based on the adaptive fusion of RGB and skeleton features to efficiently improve the accuracy of action recognition. However, the conventional action recognition algorithms based on RGB and skeleton features generally suffer from the inability to effectively utilize the complementarity of the two features and thus fail to focus on important frames in the video. Considering this, we first used the bidirectional long short-term memory network (Bi-LSTM) combined with a self-attention mechanism to extract spatial-temporal features of RGB and skeleton images. Next, we constructed an adaptive weight computing network (AWCN) and computed these adaptive weights based on the spatial features of two types of images. Finally, the extracted spatial-temporal features were fused by the adaptive weights to implement action recognition. Using Penn Action, JHMDB, and NTU RGB-D dataset, the experimental results show that our proposed algorithm effectively improves the accuracy of action recognition compared with existing methods.

Key words machine vision; action recognition; pose estimation; adaptive weight computing network; long short-term memory network; self-attention
OCIS codes 150.1135; 100.3008; 100.4996; 100.5010

1 引 言

目前，人体行为识别已在多个领域广泛应用，如安防监控、视频搜索、人机交互等，故得到了人们越来越多的关注。随着人体行为和其他领域的紧密结合，行为采集和识别所获取的信息给科学研究提供了极大的便利条件，在多个领域具有大量的应用场景。人体行为识别已经成为人工智能领域热门的研究方向之一。

在早期的文献中，文献[1]通过计算相机的运

收稿日期：2019-12-23；**修回日期**：2020-01-19；**录用日期**：2020-02-25
基金项目：国家自然科学基金（61362030，61201429）、中国博士后科学基金（2015M581720，2016M600360）、科技援疆专项计划（2017E0279）、江苏博士后科学基金（1601216C）

* **E-mail**：kongjun@jiangnan.edu.cn

动，提出改进的稠密轨迹，得到了较好的识别效果。文献[2]通过提取 RGB 图像的纹理、颜色和梯度方向等基础特征，使用基于二分类的多类 LogitBoost 分类器进行分类识别。仅使用 RGB 图像，容易受到光照和尺度变换的影响，识别结果不尽人意。文献[3]提出基于四元数三维(3D)骨骼表示的算法，准确地描述了 3D 骨骼间的几何关系，提高了行为识别的准确率。文献[4]通过将深度图序列转换为三维点云序列，有效地降低了空间尺度变化对识别准确率的影响。但是，如文献[5]和文献[6]中所述，仅使用骨骼图像或深度图像在面对相似的动作时，如喝水和打电话时，识别效果并不理想。如何高效地融合 RGB 图像和骨骼图像特征，提升特征互补性，已逐渐成为研究的新兴热点。随着脉冲耦合神经网络(P-CNN)[7]的出现，越来越多的 RGB 和骨骼特征融合方法涌现出来。文献[7]、文献[8]通过提取 RGB 图像中关节坐标附近区域的时空特征进行识别。文献[9]将相同关节的坐标点放在同一张图像中，得到了不同关节的运动轨迹图。利用卷积神经网络提取轨迹图的特征得到 PoTion 特征，然后融合 PoTion 特征、RGB 特征、光流特征进行分类。文献[10]分别对 RGB 图像和骨骼图像提取时空特征，然后融合两种时空特征进行分类。文献[11]通过提取骨骼特征、人体轮廓、RGB 图像的时空特征，融合三种时空特征进行分类。文献[12]利用长短时记忆(LSTM)网络提取 RGB 和骨骼特征后在多个尺度上进行识别并融合两者结果。文献[13]通过 VGG-19[14]关注骨骼关节附近 RGB 图像和光流图像的时空特征，并融合两种特征进行分类。上述方法在识别时都取得了较好的效果。但是在融合特征时，只是简单地融合多种时空特征，并没有考虑到多种特征之间的互补性。提取时序特征时，平等地对待每一帧，无法重点关注视频中的重点帧，视频的关键时序性不强。

综上，为了有效地利用 RGB 图像和骨骼图像之间的互补性，以及着重关注视频中的重点帧，提出联合 RGB 和骨骼特征的自适应融合的行为识别算法。首先，针对 RGB 图像和骨骼图像，通过本文提出的结合自注意力机制的双向 LSTM 提取两者时空注意力特征；然后，以 RGB 图像和骨骼图像的空间特征为输入，通过本文提出的自适应权重计算网络计算得到两种特征的融合权重；最后，基于获得的自适应权重，融合二者时空特征并通过 Softmax 进行最终分类。

2 理论基础

仅使用 RGB 图像进行分类时，分类结果容易受到 RGB 图像的光照和尺度变换等的影响。仅使用骨骼图像在面对相似的动作时，因为缺少与人体交互的物体信息[5-6]，会导致识别效果不理想，所以本文使用 RGB 图像和骨骼图像两种特征来识别动作的类别。为了提高算法的泛化性，通过 Pose Estimation[15]算法计算得到骨骼图像。使用 Pose Estimation 可以自由地提取视频中的人体骨骼图像，不会受到 Kinect 摄像头的硬件限制。RGB 图像和它对应的骨骼图像如图 1 所示，图中从左到右为打棒球、打高尔夫、跑步以及散步图像。首先利用 Inception V3 以及融合了自注意力机制的双向 LSTM 提取每一视频的 RGB 图像和骨骼图像的时空特征。然后将 RGB 图像和骨骼图像的空间特征按帧序排列，输入到自适应权重计算网络中，针对该视频的整体情况计算两种特征的自适应融合权重。最后将两种时空特征按自适应权重进行融合。算法的整体框架如图 2 所示，分为 RGB 和骨骼的时空特征提取模块、自适应权重计算模块以及特征融合分类模块。

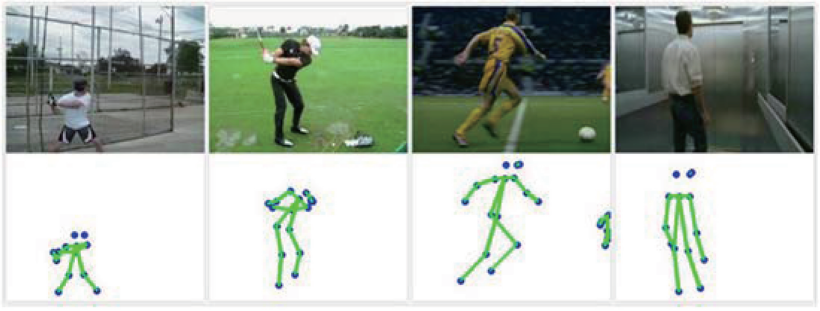

图 1 RGB 与其对应的骨骼图像

Fig. 1 RGB images and corresponding skeleton images

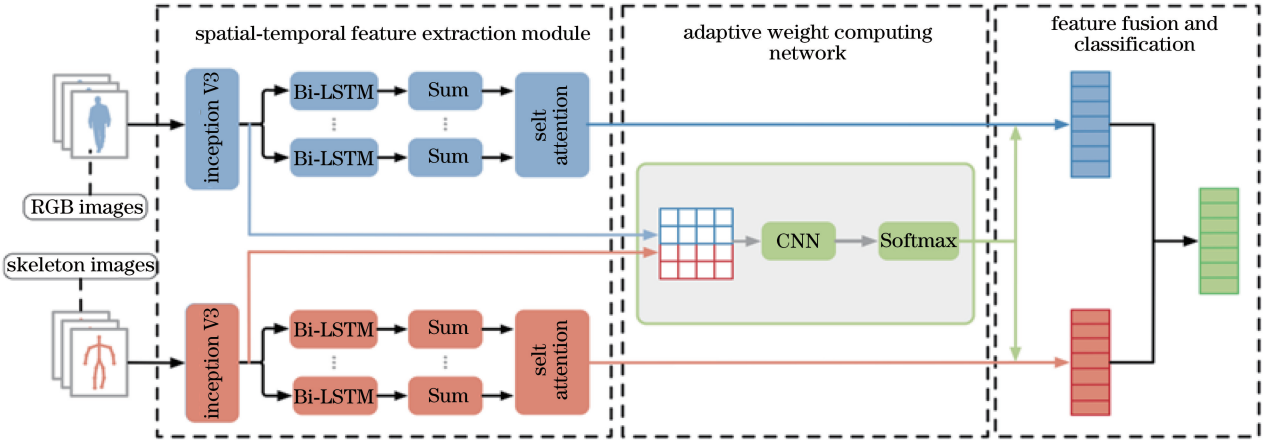

图 2 整体网络

Fig. 2 Overall network

2.1 时空特征提取模块

在空间特征提取部分，为了更加高效地提取空间特征，采用 Inception V3 网络。Inception V3 通过不同尺寸的卷积核对图像进行卷积操作，将不同的卷积层得到的特征拼接在一起，最后通过全连接层输出 $n\times1024$ 的空间特征（n 为每个视频提取的帧数）。

在时序特征提取部分，为了更加高效地提取时序特征，采用了双向 LSTM 网络。因为在对行为进行识别的时候，当前帧不仅依赖之前的视频帧，同时也依赖其之后的视频帧，故只有将两个方向的时序信息相互结合，才能提高行为识别的准确率。早期的算法都是对最后一帧两个方向的特征进行相加，不能体现不同帧对视频重要性的差异。因为不是所有帧都对识别结果有着相同的影响，有些帧对识别结果有更大的影响，而有些帧则可能对识别结果有着错误的引导作用进而导致误分类。比如，跳远之前的助跑，如果助跑的时间比跳远的时间更长，就很可能会误分类为跑步。所以，需利用自注意力机制来关注对识别结果更加重要的视频帧。通过自注意力机制对视频中的帧分配不同的权重，对识别结果的影响越大，则权重越大，对识别结果影响越小，则权重越小。

时空特征提取模块如图 3 所示。首先，通过 Inception V3 分别对 RGB 图像或者骨骼图像提取空间特征。将 Inception V3 提取的空间特征输入到两个方向的 LSTM 中，通过两个方向的 LSTM 网络提取每一帧与其前后帧的关系，输出特征向量，表达式为

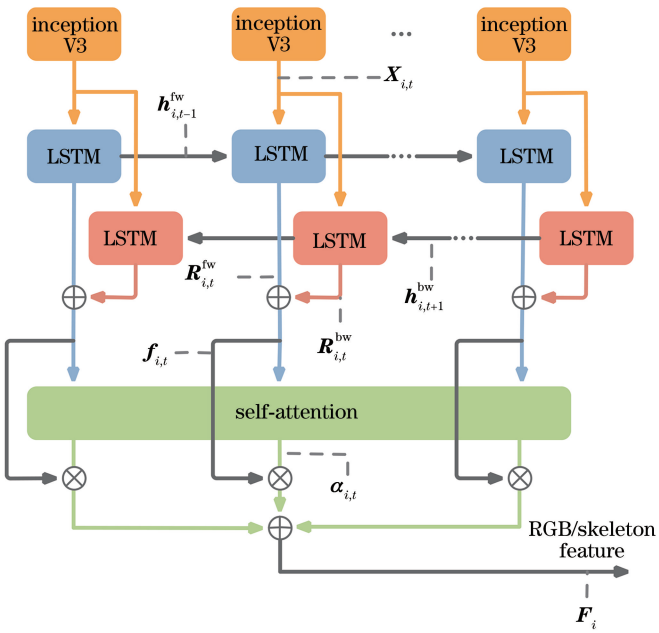

图 3 自注意力时空特征提取网络

Fig. 3 Spatial-temporal feature extracting network with self-attention

$$\boldsymbol{R}_{i,t}^{\mathrm{fw}}=\mathrm{LSTM}_{\mathrm{fw}}(\boldsymbol{h}_{i,t-1}^{\mathrm{fw}},\boldsymbol{X}_{i,t}),\tag{1}$$

$$\boldsymbol{R}_{i,t}^{\mathrm{bw}}=\mathrm{LSTM}_{\mathrm{bw}}(\boldsymbol{h}_{i,t+1}^{\mathrm{bw}},\boldsymbol{X}_{i,t}),\tag{2}$$

$$\boldsymbol{f}_{i,t}=\boldsymbol{R}_{i,t}^{\mathrm{fw}}+\boldsymbol{R}_{i,t}^{\mathrm{bw}},\tag{3}$$

式中：$\mathrm{LSTM}_{\mathrm{fw}}(\cdot)$为正向 LSTM 的计算公式；$\mathrm{LSTM}_{\mathrm{bw}}(\cdot)$为反向 LSTM 的计算公式；$t$ 表示第 t 帧；i 为第 i 个样本；$\boldsymbol{h}$ 为经过 LSTM 网络的隐藏状态向量；$\boldsymbol{X}$ 为 Inception 网络提取的空间特征；$\boldsymbol{f}_{i,t}$ 为融合的时空特征。利用本文提出的自注意力机制计算每一帧图像在分类时所占的比重，如：在“跳远”这一动作中，自注意力机制会分配给“起跳”这一动作更多的权重，而分配给“起跳”之外动作如“助跑”的权重则会较少。在分类时，网络会更加关注权重较高的视频帧的时空特征。通过对每一帧分配不同的权重，使网络能够自动地关注视频序列中的关键帧。利用不同的权重突出视频序列中更为重要的时序特征，使时空特征提取网络能够表现出更好的性能。最后，融合视频序列中各帧的时空特征，得到每个视频的 RGB 或者骨骼的时空注意力特征，表达式为

$$\boldsymbol{F}_i=\sum_{t=1}^{n}\boldsymbol{\alpha}_{i,t}\times\boldsymbol{f}_{i,t},\tag{4}$$

$$\boldsymbol{\alpha}_{i,t}=\frac{\exp(\boldsymbol{u}_{i,t}\times\boldsymbol{v})}{\sum_{t=1}^{n}\exp(\boldsymbol{u}_{i,t}\times\boldsymbol{v})},\tag{5}$$

$$\boldsymbol{u}_{i,t}=\mathrm{sigmoid}(\boldsymbol{W}\times\boldsymbol{f}_{i,t}+\boldsymbol{b}),\tag{6}$$

式中：$\boldsymbol{W}$ 为初始化的维度转换矩阵 $\boldsymbol{R}^{1024\times1}$；$\boldsymbol{b}$ 为初始化的偏置矩阵 $\boldsymbol{R}^{1\times1}$；$\boldsymbol{v}$ 为初始化的偏置矩阵 $\boldsymbol{R}^{1\times1}$；$n$ 为视频提取出的总帧数。

2.2 自适应权重计算网络

本研究基于 RGB 特征和骨骼特征提出自适应权重计算网络。在对视频进行分类时，应该针对每一视频使用适合该视频的特征进行分类，如：打高尔夫和打棒球利用 RGB 特征分类的准确率更高，因此，在特征融合时 RGB 特征应该在最终的融合特征中占的比重更大；跑步和散步利用骨骼特征分类的准确率更高，因此，在特征融合时骨骼特征应该在最终的融合特征中占的比重更大。不同的样本面对的情况不同，针对所有样本使用传统固定的融合比例不能有效地解决这个问题。为此，提出了自适应权重计算网络，如图 4 所示。

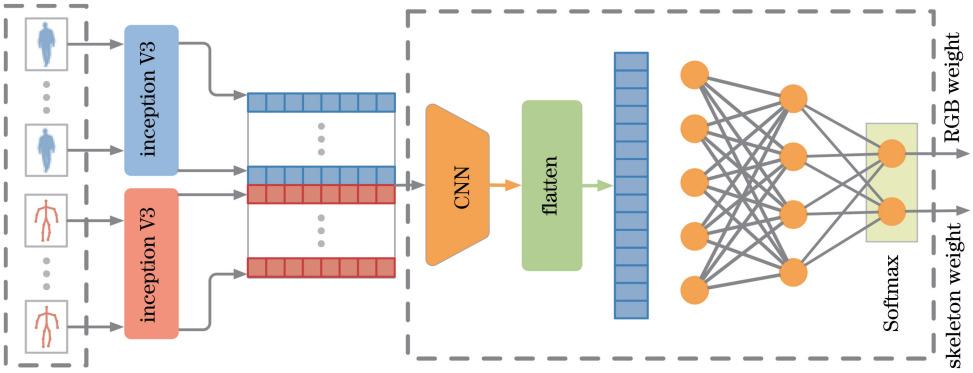

图 4 自适应权重计算网络

Fig. 4 Adaptive weight computing network

对每个样本提取 15 帧 RGB 图像，并利用 Pose Estimation 计算骨骼图像。利用 Inception V3 网络对 RGB 图像以及骨骼图像提取空间特征，输出维度为 1×1024 的特征向量。RGB 和骨骼图像的特征向量分别按帧序排列，得到 15×1024 的特征向量，再将 RGB 和骨骼图像的特征向量拼接得到 30×1024 的特征向量，并将其看作一幅图像，这个图像代表了视频的整体特征。通过分析这个 30×1024 的特征向量，可以计算出 RGB 特征和骨骼特征的融合权重。所以使用卷积神经网络(CNN)从这个特征向量中提取特征，然后连接全连接网络并使用 Softmax 作为激活函数，这样可以得到一个二维的向量。这个向量就是 RGB 特征和骨骼特征的自适应融合权重。通过对每个视频样本计算适合该样本的自适应权重融合的两种时空注意力特征，可以提升整个数据集的识别精度。自适应权重融合的表达式为

$$\tilde{\boldsymbol{\omega}}_i=H\{\mathrm{Conv}\{G[\boldsymbol{I}(S_{i,1}^{\mathrm{RGB}}),\boldsymbol{I}(S_{i,2}^{\mathrm{RGB}}),\cdots,\boldsymbol{I}(S_{i,n}^{\mathrm{RGB}}),\boldsymbol{I}(S_{i,1}^{\mathrm{Skeleton}}),\boldsymbol{I}(S_{i,2}^{\mathrm{Skeleton}}),\cdots,\boldsymbol{I}(S_{i,n}^{\mathrm{Skeleton}})]\}\},\tag{7}$$

式中：$\tilde{\boldsymbol{\omega}}$ 为获得的融合比例矩阵；$S_{i,n}^{\mathrm{RGB}}$ 为第 i 个视频的第 n 帧 RGB 图像；$S_{i,n}^{\mathrm{Skeleton}}$ 为第 i 个视频的第 n 帧

骨骼图像；$\boldsymbol{I}(\cdot)$为提取视频帧的 Inception V3 空间特征；$G(\cdot)$为将 RGB 图像和骨骼图像的空间特征按帧序排列并进行拼接操作；Conv$(\cdot)$为对通过拼接获得的特征向量作卷积操作；$H(\cdot)$为非线性激活函数 Softmax。自适应权重计算网络能够根据每一视频特征计算适合该样本的融合权重，而不是针对所有样本设置同样的融合权重，故能够有效地利用两种特征之间的互补性。

2.3 特征融合分类模块

图 5 给出本文采用的特征融合分类模块。该模块将每个视频的 RGB 和骨骼时空特征以两者的融合比例进行融合，获得融合特征 $\boldsymbol{T}_i$，随后通过 Softmax 作激活函数对该视频进行分类。融合特征可表示为

$$\boldsymbol{T}_i=\tilde{\boldsymbol{\omega}}_i^{\mathrm{RGB}}\times\boldsymbol{F}_i^{\mathrm{RGB}}+\tilde{\boldsymbol{\omega}}_i^{\mathrm{Skeleton}}\times\boldsymbol{F}_i^{\mathrm{Skeleton}},\tag{8}$$

$$P_j(\boldsymbol{T}_i)=\frac{\exp(\boldsymbol{T}_i^j)}{\sum_{j=1}^{K}\exp(\boldsymbol{T}_i^j)},\tag{9}$$

式中：$\tilde{\boldsymbol{\omega}}_i^{\mathrm{RGB}}$ 和 $\tilde{\boldsymbol{\omega}}_i^{\mathrm{Skeleton}}$ 是通过(7)式得到的第 i 个视频的 RGB 和骨骼的时空注意力特征的融合比例；$\boldsymbol{F}_i^{\mathrm{RGB}}$ 和 $\boldsymbol{F}_i^{\mathrm{Skeleton}}$ 是通过(4)式得到的第 i 个视频的 RGB 和骨骼的时空注意力特征；$P_j(\boldsymbol{T}_i)$为特征 T 属于第 j 类的概率值；K 表示该数据集总共有 K 个类别。本研究针对每个样本计算其 RGB 和骨骼特征的自适应融合权重，在有效利用两种特征之间的互补性的同时提高了识别精度。此外，本文采用的骨骼图像由 Pose Estimation 计算得到，不受 Kinect 摄像头的硬件限制，极大地提高了算法的泛化性。

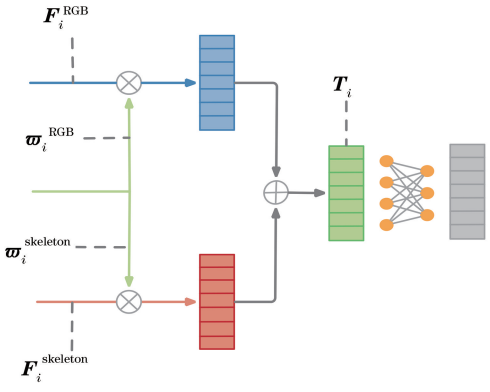

图 5 特征融合分类模块

Fig. 5 Feature fusion and classification

3 实验研究

3.1 实验数据集

JHMDB 数据集由 21 类共 928 个短视频组成。数据集的视频都是从互联网采集而来，视频的拍摄角度和人物着装都不相同。人物在视频图像中不完整，给行为识别也带来了很大的挑战。数据集的测试方式有三种，为了便于与其他文献作对比，取三种测试方式结果的平均值。

Penn Action 数据集包含了 15 类共 2326 个视频样本。数据集的样本同样是从互联网采集而来，所有的视频样本都包含了较大的尺度变化。视频样本的人物着装、背景、拍摄角度都不一样。按照数据集的测试方法，选择一半的数据作训练集，另一半的数据作测试集。

NTU RGB-D 数据集包含了 60 类共 56880 个视频样本。数据集的样本是通过不同的人物在不同的场景下采集而来。数据集的测试方法分为两种：Cross-Subject 和 Cross-View。Cross-Subject 选择 20 个人物的数据作训练集，20 个人物的数据作测试集。Cross-View 选择 2 个视角的数据作训练集，1 个视角的数据作测试集。

3.2 实验设置

本实验使用 Python 语言在 GPU 加速环境下进行实验，采用 Keras 深度学习框架，计算机配置为 Ubuntu16.04 系统、64 GB 内存、2×GTX1080 11 G 显存。在使用 Inception V3 时屏蔽了最后的 5 层，最后的 5 层只用来提取特征。表 1 中列举了本实验的参数设置。

表 1 实验参数

Table 1 Experimental parameters

Parameter	Value
Loss function	Categorical cross entropy
Optimizer	Adam
Learning rate	0.0001
Batch_size	32
Number of epoch	150

3.3 精度影响因素分析

3.3.1 自注意力

综上所述，并不是所有的帧都对行为识别有益，与类别无关的行为会扰乱行为识别的最终结果。通过使用自注意力机制，网络会自动给每一帧分配相应的权重，越有利的帧其权重越高，学习到的时序特征也会更有效，最终改善了行为识别的精度。在 Penn Action 和 JHMDB 数据集上，对使用自注意力机制前后的 RGB 特征、骨骼特征以及融合特征进行了对比实验。从表 2 和表 3 可以看出，使用自注意力后相较于未使用自注意力时的识别精度提升了 1.5%左右。在使用自注意力机制后，仅使用 RGB 特征、仅使用骨骼特征以及使用两者融合特征的识

别效果都有了明显的提升，由此说明网络能够更有效地提取视频序列的时序特征。

表 2 在 Penn Action 数据集下使用与未使用自注意力识别精度

Table 2 Accuracy with and without self-attention on Penn Action dataset unit：%

Attention	RGB	Skeleton	Fusion
Without attention	90.3	83.8	92.8
With attention	92.1	85.2	94.3

表 3 在 JHMDB 数据集下使用与未使用自注意力识别精度

Table 3 Accuracy with and without self-attention on JHMDB dataset unit：%

Attention	RGB	Skeleton	Fusion
Without attention	69.2	61.9	72.9
With attention	71.3	63.7	74.8

3.3.2 固定权重和自适应权重

AWCN 能够按照视频的整体特征来调整 RGB 和骨骼特征的融合比例。为了进一步验证采用 AWCN 的有效性，在两个数据集上进行了关于固定权重和自适应权重的对比实验（未使用自注意力机制）。

如图 6 所示，设置了几组固定权重组合和自适应权重的实验。横轴代表权重组合，纵轴代表识别精度。权重组合分别有(0.8,0.2)，(0.6,0.4)，(0.5,0.5)，(0.4,0.6)，(0.2,0.8)。每个权重组合中的第一个数据为 RGB 特征的权重，第二个数据为骨骼特征的权重。从表中可以看出两个数据集的识别精度在(0.8,0.2)和(0.5,0.5)之间上升。然后随着 RGB 特征的权重的减小，精度开始降低。这是因为识别结果开始倾向于骨骼特征，而骨骼特征的识别精度比 RGB 特征的识别精度低。所以，如果增加骨骼特征的融合比率，精度就会下降。比较最高精度权重组合(0.5,0.5)和自适应权重（如图 6 中虚线框所示）的识别结果，可以发现自适应权重的融合结果比(0.5,0.5)的结果更好。这是因为自适应权重计算网络可以针对每个视频计算适合该视频的融合比例，且自适应权重不需要手动调整融合比例，在提高效率的同时提高了特征间的互补性。

3.3.3 不同数据集

本文采用的 AWCN 与现有方法的比较如表 4 所示。可以看出 AWCN 在 Penn Action 数据集上有着良好的表现。C3D 只关注了 RGB 特征，并没有关注骨骼特征，所以精度相较于本文识别精度较低。JDD 利用 C3D 提取关节点处的 RGB 特征并予以分类，如果关节点位置预测错误，则会造成识别错误，而本文方法对 RGB 特征和骨骼特征分别同时提取特征，互不影响。AOG 使用三个尺度提取关节处的特征，同样有着类似于 JDD 的问题。IDT-FV+Pose 只是简单地使用固定权重融合 IDT-FV 和 Pose（骨骼）特征，未考虑到特征间的互补性，所以精度比本文方法偏低。TSN 融合了 RGB 和光流特征，获取光流图像时会耗费大量的时间。DPI 和 MMTSN 在融合 RGB 和骨骼特征时，未考虑到特征间的互补性，不能针对不同的视频选择适合视频的特征进行分类。

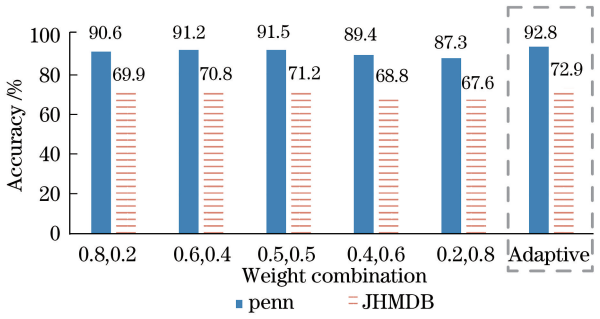

图 6 不同权重组合的识别精度

Fig. 6 Accuracy of different weight combinations

表 4 Penn Action 数据集上 AWCN 与其他模型算法比较

Table 4 Comparison of AWCN and other algorithms on Penn Action dataset unit：%

Algorithm	Accuracy
AOG-Fine[16]	73.4
STIP-HoG+HoG[17]	82.8
AOG-All[16]	85.5
C3D[18]	86.0
JDD[19]	87.4
MMTSN-RGB+Pose[20]	91.67
IDT-FV[19]	92.0
IDT-FV+Pose[19]	92.9
TSN[21]	93.8
DPI+att-DTI[22]	93.9
DPI+att-DTIs[22]	95.8
AWCN (Ours)	92.8
AWCN+self-attention (Ours)	94.3

从表 5 中可以看出，本文 AWCN 在 JHMDB 数据集上同样有着良好表现。P-CNN、TS R-CNN、MR-TS R-CNN 都是从关节附近提取 RGB 和光流特征，所以错误的关节坐标会对识别造成很大的影响，同时光流图像的计算会耗费大量的时间。FAT 通过骨骼坐标将图像分为 4 个尺度，利用固定的权重融合 RGB 和光流特征，不能有效地利用两种特征

的互补性，所以精度比 AWCN 低。GoogLeNet＋iTF 通过提取视频的光流图像进行行为识别，仅使用单个模态数据，识别精度较低，同时光流图像的获取耗费了大量的时间。MMTSN 在使用 RGB 和 Pose 特征融合时，提取关键时序特征的能力不强。同时，通过固定的权重融合两种特征，未能有效地利用特征间的互补性。

表 5 JHMDB 数据集上 AWCN 与其他模型算法比较

Table 5 Comparison of AWCN and other algorithms on JHMDB dataset unit：%

Algorithm	Accuracy
P-CNN[7]	61.1
FAT[23]	62.5
MMTSN-RGB＋Pose[20]	62.86
STAR-Net[24]	64.3
IDT-FV[19]	65.9
TS R-CNN[23]	70.5
MR-TS R-CNN[23]	71.1
GoogLeNet＋iTF[25]	74.5
AWCN (Ours)	72.9
AWCN＋self-attention (Ours)	74.8

从表 6 中可以看出，本文 AWCN 在 NTU RGB-D 数据集上同样有着良好表现。VA-LSTM、STA-LSTM 和 CSTA-CNN 仅使用骨骼特征进行行为识别，这几类方法在面对相似的动作时，缺少 RGB 特征，会对识别造成很大的影响。Two-Stream CNN 利用拼接的方式融合骨骼空间和运动特征，不能有效地利用两种特征的互补性。ST-GCN在提取时空特征时，提取关键时序特征能力不强，容易被无关动作干扰造成误分类，同样在面对相似的动作时，会因为 RGB 特征的缺失，对识别造成很大的干扰。HCN 提取视频帧中的每个人物，分别识别每个人的动作，未能有效地利用视频中多个人物动作的交互性。

表 6 NTU RGB-D 数据集上 AWCN 与其他模型算法比较

Table 6 Comparison of AWCN and other algorithms on NTU RGB-D dataset unit：%

Algorithm	CS	CV
STA-LSTM[26]	73.4	81.2
VA-LSTM[27]	79.4	87.6
ST-GCN[28]	81.5	88.3
Two-Stream CNN[29]	83.2	89.3
CSTA-CNN[30]	84.9	89.9
HCN[31]	86.5	91.9
SR-TSL[32]	84.8	92.4
AWCN (Ours)	85.6	88.9
AWCN＋self-attention (Ours)	87.3	90.1

3.3.4 仅使用骨骼特征与使用融合特征

为了验证相似的动作若仅使用骨骼特征会导致识别效果不理想这一结论，抽取了 Penn Action 数据集中的 Golf 和 Baseball Swing 两种动作的样本，对仅使用骨骼特征与使用 RGB 和骨骼融合特征的识别结果进行对比实验。从图 7 左图中观察可以发现仅使用骨骼特征时，两种动作的误分类率较高；如图 7 中右图所示，使用融合特征后，两种动作的识别准确率有了显著的提升。由此可以得出，若仅使用骨骼特征，相似的动作会因为缺少与人体交互的物体特征而导致识别效果不理想。

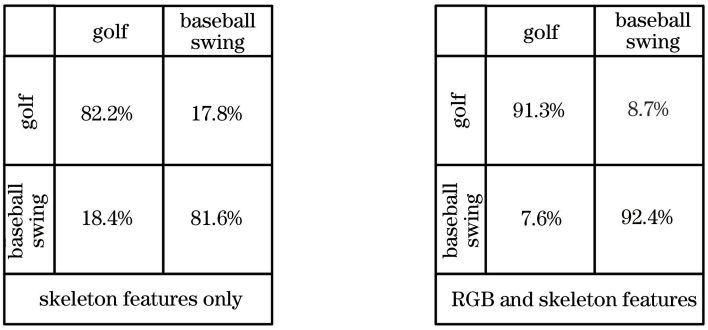

图 7 仅使用骨骼特征和使用融合特征的识别结果

Fig. 7 Recognition results of using skeleton features only and fusion features

3.4 实验结果可视化

3.4.1 自注意力机制可视化

图 8 和图 9 分别为 Penn Action 中 Golf 和 Baseball Swing 的自注意力机制的可视化效果图。本文提出的时空特征提取网络从每个视频序列中提取 15 帧图像作为输入。为了观察方便，本文从 15 帧视频序列中抽取 7 帧。图中骨骼图像和 RGB 图像下方的数字表示根据自注意力机制，分别计算得到的对应

帧相对于整体视频序列在分类时所占的比重。图中加粗边框表示对识别更加重要的视频帧。通过观察可以得出在击打高尔夫球的第 2、3、4、5、6 列的权重相对较高，在击打棒球的第 3、4、5、6、7 列的权重相对较高。而这几帧图像正是击打高尔夫球和棒球的关键动作。若未使用自注意力机制，则视频序列中所有帧的权重都相同，时空特征提取网络不能自动地关注视频序列中的关键帧。通过使用自注意力机制，网络能够给关键的视频帧分配更高的权重，从而提升网络的时空特征提取能力，进而提升识别精度。

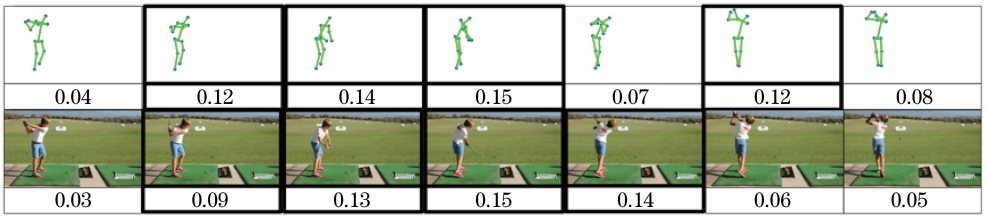

图 8 Golf 骨骼图像和 RGB 图像自注意力可视化效果图

Fig. 8 Visualization of self-attention on skeleton and RGB images of Golf

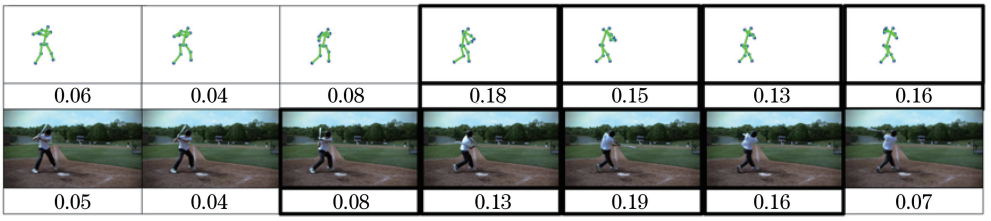

图 9 Baseball swing 骨骼图像和 RGB 图像自注意力可视化效果图

Fig. 9 Visualization of self-attention on skeleton and RGB images of Baseball swing

3.4.2 自适应权重可视化

图 10 中从上至下分别为 Penn Action 中 Golf、Baseball Swing、Walk 以及 Run 的自适应权重计算网络的可视化效果图。为了观察方便，本文从 15 帧视频序列中抽取 7 帧。骨骼图像和 RGB 图像左方的数字表示根据自适应权重计算网络，计算得到的两种特征在最终的融合特征中所占的权重。图中加粗边框为对融合特征较为重要的特征。通过观察，可以得出 Golf 和 Baseball Swing 的骨骼图像，缺少与人体交互的物体的特征。而区分 Golf 和 Baseball Swing 两种动作更需要 RGB 图像中的背景以及与人体交互的物体即球棒等信息，所以在融合特征中 RGB 特征所占的比重相对较高。Walk 和 Run 的 RGB 图像中光照条件不理想，但是骨骼图像不受光照的影响，故这两幅图像可以通过骨骼特征更好地区分两种动作，所以在融合特征中骨骼

图 10 Golf、Baseball swing、Walk 和 Run 的自适应权重可视化效果图

Fig. 10 Visualization of adaptive weight of Golf, Baseball swing, Walk and Run

特征所占的比重相对较高。若未使用自适应权重计算网络，则所有的视频样本都使用相同的融合权重，网络不能有效地利用特征之间的互补性。通过使用自适应权重计算网络，网络能够自动地针对每一个视频样本计算适合该样本的两种特征融合权重，从而有效地利用了特征之间的互补性，进而提升网络的识别能力。

4 结 论

本文为了关注对识别结果影响更大的视频帧，提出了结合自注意力机制的双向 LSTM，有效地提高了识别精度。利用 RGB 和骨骼图像的空间特征训练了自适应权重计算网络，以针对每一视频计算两种特征的自适应融合权重。基于自适应融合权重，能够融合 RGB 和骨骼特征，得到相比于固定权重更好的识别结果。在 Penn Action、JHMDB 和 NTU RGB-D 数据集上的实验能够证明本文方法的优越性。

参 考 文 献

[1] Wang H, Schmid C. Action recognition with improved trajectories[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 3551-3558.

[2] Li C J, Liu Y. Abnormal driving behavior detection based on covariance manifold and LogitBoost[J]. Laser & Optoelectronics Progress, 2018, 55(11): 111503.
李此君, 刘云鹏. 基于协方差流形和 LogitBoost 的异常驾驶行为识别方法[J]. 激光与光电子学进展, 2018, 55(11): 111503.

[3] Xu H Y, Kong J, Jiang M. Human action recognition based on quaternion 3D skeleton representation [J]. Laser & Optoelectronics Progress, 2018, 55(2): 021002.
徐海洋, 孔军, 蒋敏. 基于四元数 3D 骨骼表示的人体行为识别[J]. 激光与光电子学进展, 2018, 55(2): 021002.

[4] Xu H Y, Kong J, Jiang M, et al. Action recognition based on histogram of spatio-temporal oriented principal components[J]. Laser & Optoelectronics Progress, 2018, 55(6): 061009.
徐海洋, 孔军, 蒋敏, 等. 基于时空方向主成分直方图的人体行为识别[J]. 激光与光电子学进展, 2018, 55(6): 061009.

[5] Hu J F, Zheng W S, Lai J H, et al. Jointly learning heterogeneous features for RGB-D activity recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 5344-5352.

[6] Liu T S, Kong J, Jiang M. RGB-D action recognition using multimodal correlative representation learning model[J]. IEEE Sensors Journal, 2019, 19(5): 1862-1872.

[7] Chéron G, Laptev I, Cordelia. P-CNN: pose-based CNN features for action recognition[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 3218-3226.

[8] Baradel F, Wolf C, Mille J. Pose-conditioned spatio-temporal attention for human action recognition[EB/OL]. (2017-08-07)[2019-12-18]. https://arxiv.org/abs/1703.10106.

[9] Choutas V, Weinzaepfel P, Revaud J, et al. PoTion: pose MoTion representation for action recognition[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7024-7033.

[10] Luvizon D C, Picard D, Tabia H. 2D/3D pose estimation and action recognition using multitask deep learning [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5137-5146.

[11] El-Ghaish H, Hussien M E, Shoukry A, et al. Human action recognition based on integrating body pose, part shape, and motion[J]. IEEE Access, 2018, 6: 49040-49055.

[12] Luvizon D C, Tabia H, Picard D. Multi-task deep learning for real-time 3D human pose estimation and action recognition[EB/OL]. (2019-12-15)[2019-12-18]. https://arxiv.org/abs/1912.08077.

[13] Cai Z, Neher H, Vats K, et al. Temporal Hockey Action Recognition via Pose and Optical Flows[EB/OL]. (2018-12-22)[2019-12-18]. https://arxiv.org/pdf/1812.09533.pdf.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2019-12-18]. https://arxiv.org/abs/1409.1556.

[15] Cao Z, Simon T, Wei S, et al. Realtime multi-

person 2D pose estimation using part affinity fields [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1302-1310.

[16] Nie B X, Xiong C M, Zhu S C. Joint action recognition and pose estimation from video [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1293-1301.

[17] Zhang W, Zhu M, Derpanis K G. From actemes to action: a strongly-supervised representation for detailed action understanding[C]//Proceedings of the IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 2248-2255.

[18] Cao C, Zhang Y, Zhang C, et al. Action recognition with joints-pooled 3D deep convolutional descriptors [C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: ACM, 2016: 3324 - 3330.

[19] Iqbal U, Garbade M, Gall J. Pose for action-action for pose [C] // 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), May 30 - June 3, 2017, Washington, DC, USA. New York: IEEE Press, 2017: 438-445.

[20] Khalid M U, Yu J. Multi-modal three-stream network for action recognition [C] // 2018 24th International Conference on Pattern Recognition (ICPR), August 20-24, 2018, Beijing, China. New York: IEEE Press, 2018: 3210-3215.

[21] Du W B, Wang Y L, Qiao Y. RPAN: an end-to-end recurrent pose-attention network for action recognition in videos [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 3745-3754.

[22] Liu M Y, Meng F Y, Chen C, et al. Joint dynamic pose image and space time reversal for human action recognition from videos [C] // Proceedings of the AAAI Conference on Artificial Intelligence, Reston, VA: AIAA, 2019, 33: 8762-8769.

[23] Peng X, Schmid C. Multi-region two-stream R-CNN for action detection [C] // European Conference on Computer Vision. Cham: Springer , 2016: 744-759.

[24] McNally W, Wong A, McPhee J. STAR-net: action recognition using spatio-temporal activation reprojection[C]//2019 16th Conference on Computer and Robot Vision (CRV), May 29-31, 2019, Kingston, QC, Canada. New York: IEEE Press, 2019: 49-56.

[25] Javidani A, Mahmoudi-Aznaveh A. Learning representative temporal features for action recognition [EB/OL]. (2018-02-19) [2019-12-18]. https: // arxiv.org/abs/1802.06724.

[26] Song S J, Lan C L, Xing J L, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data [EB/OL]. (2016-11-18) [2019-12-18]. https: // arxiv. org/abs/1611.06067.

[27] Zhang P F, Lan C L, Xing J L, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data [J]. 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2136-2145.

[28] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[EB/OL]. (2018-01-23)[2019-12-18]. https: //arxiv.org/abs/1801.07455.

[29] Li C, Zhong Q Y, Xie D, et al. Skeleton-based action recognition with convolutional neural networks [C] // 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), July 10-14, 2017, Hong Kong, China. New York: IEEE Press, 2017: 597-600.

[30] Wang J Y. Learning Coupled Spatial-temporal Attention for Skeleton-based Action Recognition[EB/OL]. (2019-09-23) [2019-12-18]. https: // arxiv.org/abs/1909.10214.

[31] Li C, Zhong Q Y, Xie D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation [EB/OL]. (2018-04-17) [2019-12-18]. https: //arxiv.org/abs/1804.06055.

[32] Si C Y, Jing Y, Wang W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning network [J]. Pattern Recognition, 2020, 107: 107511.