Computer Science and Application 计算机科学与应用, 2021, 11(2), 451-460
Published Online February 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.112045

Hans汉斯

融合多模态特征的多流行为识别网络

张彬彬，江朝晖，李君君

合肥工业大学计算机与信息学院，安徽 合肥
Email: 2055781665@qq.com, 827894689@qq.com, 17201577500@163.com

收稿日期：2021年1月23日；录用日期：2021年2月18日；发布日期：2021年2月25日

摘　要

针对当前行为识别网络抗干扰能力不足和单一特征难以鲁棒性的表达行为的问题，本文提出了一种融合多模态特征的多流行为识别网络模型。首先，利用三维神经网络来提取RGB视频帧的表观特征和光流帧的运动特征，并利用注意力机制学习重要信息的权重。同时，本文引入了一个姿态网络来建模人体姿态序列的时空特征，弥补表观特征和运动特征对行为表达能力的不足。最后通过对三种特征的学习来实现行为识别。本文在JHMDB数据集上进行实验验证，结果表明我们的方法优于当前大多数先进的方法。

关键词

行为识别，注意力机制，姿态序列，3D卷积，姿态网络

Multi-Stream Action Recognition Network Fusing Multi-Modal Features

Binbin Zhang, Chaohui Jiang, Junjun Li

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui
Email: 2055781665@qq.com, 827894689@qq.com, 17201577500@163.com

Received: Jan. 23rd, 2021; accepted: Feb. 18th, 2021; published: Feb. 25th, 2021

Abstract

Aiming at the problems of insufficient anti-interference ability of current action recognition networks and the difficulty of expressing action robustly with a single feature, this paper proposes a multi-modality feature fusion multi-behavior recognition network model. First, use a three-dimensional neural network to extract the apparent features of RGB video frames and the motion features of optical flow frames, and the attention mechanism is used to learn the weight of important information. At the same time, a pose network is introduced to model the spatial and temporal

文章引用: 张彬彬, 江朝晖, 李君君. 融合多模态特征的多流行为识别网络[J]. 计算机科学与应用, 2021, 11(2): 451-460. DOI: 10.12677/csa.2021.112045

features of human posture sequence, which makes up for the deficiency of apparent features and motion features in the expression ability of action. Finally, action recognition is realized by learning the three features. Experimental verification on JHMDB dataset shows that our method is superior to most of the current advanced methods.

Keywords

Action Recognition, Attention, Posture Sequence, 3D Convolution, Pose Network

1. 引言

近些年，随着深度学习理论的发展和完善，基于视频的人体行为识别因具有很大挑战性，吸引了越来越多的研究者专注于该领域。行为识别的目的是对一段视频或者图片序列进行分析，提取时空维度的特征，分析行为模式，识别出人在其中执行的动作。相关技术在生活的各个领域得到了广泛应用。但现有的行为识别系统仍有很大的局限性且网络容易过拟合。在真实视频场景中，背景和光照变化会造成冗余干扰，动作本身复杂多变且肢体相互遮挡，这些都使得如何鲁棒性的表达行为特征成为了一个值得深究的问题。因此，本文的研究目标是设计强泛化能力的行为识别网络，识别行为。

早期的研究主要是用卷积神经网络学习多帧 RGB 图片的表观特征来建模视频行为[1]。Karen 等人[2]开创性的提出双流网络，利用两个 2D ConvNets 分别处理单帧 RGB 图片和堆叠的密集光流帧，这种结合表观特征和运动特征的方法取得了不错的效果。但是双流 2D ConvNets 能学习到的空间信息和时间尺度的信息非常有限[3]。Tran D 等人[4]在上述工作的基础上提出了 3D ConvNets 来处理视频，用时序维度卷积核来提取视频的时序特征，性能得到很大提高。J Carreira 等人[5]结合以上各个模型的优点，提出了双流 3D ConvNets 模型，同时提取连续视频帧的表观特征和连续光流帧的运动特征并取得了不小的进步。以上这些这些方法对如何有效利用表观特征和运动特征的探索，证明了融合多模态特征是可行的且是有益的，但局限性也很明显。当视频中存在强光和大的场景变化，外观特征包含的语义信息将减少，光流对动态信息的表达也将不足。由于开源算法[6]的成熟，提取视频中人体的动态姿态序列变的可行，与外观和光流相比，人体姿态序列的轨迹对光照和场景变化具有很强的鲁棒性，其跨时序的动态变化为动作识别提供了重要的信息。因此，我们考虑在多模态信息的融合学习上加入人体姿态信息并对其跨时空维度建模来解决当前动作识别系统存在的抗干扰能力差的问题。

此外，考虑到在建模整个连续帧的表观特征和运动特征的时候，冗余信息的干扰对动作识别性能的影响会累积放大，我们引入空间注意力机制来抑制无关信息，通过学习帧级关键信息的权重来关注动作发生的区域。

与现有动作识别方法相比，本文有以下几点贡献：

1) 本文提出了一个多流网络架构，同时计算和整合三种不同模态的信息：RGB、光流、姿态序列。充分利用了不同特征之间的互补性和差异性。

2) 本文引入空间注意力机制，对动作的发生区域计算注意力得分，产生该区域增强的权重，以此来增强该区域特征的显著性，抑制杂乱的无关冗余信息，从而提高行为识别的效率和精度。

3) 为了捕捉姿态序列的空间结构特征和时序动态特征，我们用一个姿态网络对其进行跨时空建模，显著提高了视频行为识别对噪声干扰的鲁棒性。

4) 我们的方法在 JHMDB 数据集上进行试验验证，取得了很好的效果，识别精度优于当前大多数先进的方法。

2. 相关工作

行为识别的目的是利用算法对一段视频或者图像序列进行分析，识别出人在其中执行的动作，并给出分类分数。行为通常是空间上相似时序上变化连续的一系列图片帧的连接，其特有的时空信息，对行为模式的分析至关重要。如何利用好时空维度特征，是行为识别领域的重要研究课题，同时为行为识别指明了方向。

早期的行为识别主要依赖于一些传统的特征提取算法。通过提取视频帧的局部关键点的特征来描述行为。I. Laptev 等人[7]将空间和时间维度上发生明显变化的位置选作视频帧中的关键兴趣点，他们认为越是发生变化的数据越是包含行为的显著特征。这种方法是对空间关键点方法的延伸，本质上是将二维 Harris 角点检测延伸到三维。Scovanner 等人[8]利用相同的思路将二维空间描述子 SIFT 扩展到三维，通过统计时空关键点周围的 HOG 来描述特征。时空关键点的特征是对局部位置信息的描述，但是这种方法存在检测到的关键点过于稀疏的问题。于是，有些学者提出通过追踪视频帧沿时间轴的变化来描述行为特征，这也是轨迹法的核心思想。H. Wang 等人[9]通过对多尺度下的视频帧进行密集采样，然后利用光流场对采样点的运动轨迹进行追踪。除此之外，H. Wang 等人[10]进一步对视频帧沿轨迹的方向在特征点周围提取 HOG、HOF、MBH、trajectory 四种特征，并利用费希尔向量编码特征，然后利用编码后的高阶特征对 SVM 分类器进行训练。然而，这些利用浅层描述符来表征视频的方法都受限于局部特征，对复杂动作的泛化表征能力很差[11]。

随着深度学习理论的发展，基于卷积神经网络的方法不再依赖手工特征，并在行为特征表征上远胜于传统方法。相比于图片，视频多出了时序维度，Karen 等人[2]提出双流模型，在空间上学习单帧 RGB 特征，在时间上提取堆叠的光流帧的特征，使得行为识别的精度得到了明显提高。Donahue J 等人[12]提出 LSTM 和 CNN 结合进行行为识别的方法。具体而言，就是利用预训练的 CNN 网络提取 RGB 帧的表观特征，然后用 LSTM 对长序列特征的学习能力表达时间维度特征。Tran D 等人[4]通过给二维卷积增加一个时序卷积核来提取视频的时序信息，兼具了性能和训练速度的优势。Quo Vadis 等人[5]提出了双流 3D ConvNets 模型，分别提取视频帧的表观特征和光流帧的运动特征并融合，取得了优于前人的成果。

基于图的模型因其对图结构化的数据的高效表达引起了广泛的关注[13]。S. Qi 等人[14]将图模型用于解决图片和视频中检测和识别人物交互的任务。Simonovsky 等人[15]首次将图卷积用于点云分类任务。Y. Seo [16]等人将图卷积和 LSTM 相结合，同时利用了两者的优势，实现了对图结构序列时空特征的学习。S. Yan [17]等人提出了一种时空图卷积模型用于行为识别，利用图卷积运算学习空间结构特征，用时序卷积学习时序特征。注意力机制最早是为解决 NLP 任务提出的。Zhenyang Li 等人[18]和 Shikhar 等人[19]的工作研究了通过 LSTM 模型整合注意力机制的方法用于行为识别。Si 等人[29]提出了注意力增强的图卷积 LSTM 网络用于行为识别。受以上这些工作的启发，本文用一个多流网络模型来融合多模态特征并利用注意力机制来去除冗余信息，实验结果验证了本文方法的有效性。

3. 模型框架

本文的行为识别网络架构如图 1 所示，接下来我们将对模型框架作详细的介绍。本文首先利用光流算法[28]计算得到原始视频帧的光流图，以此来描述视频的动态变化。利用开源视频姿态关键点检测算法

[6]检测每个视频帧的人体关节点的二维坐标位置，跨时间序列的人体关节点的二维坐标自然的表示了动态骨架的形态，然后通过分析其运动模式来识别视频中的人体行为所属的类别。为此，本文设计了一个三流网络来学习和整合多模态特征。首先，我们将原始视频帧和提取的光流图送入 I3D 卷积网络中，提取表观特征和运动特征。为了抑制不必要的冗余信息，本文在 I3D 网络中引入了空间注意力机制，提取自 I3D 网络的特征在经过注意力机制学习后得到权重增强的新特征，并将其送入全连接层学习更高语义特征。然后，对于提取的姿态序列，本文用姿态网络建模其空间结构特征和长时序依赖。最后通过对多模态特征的学习，进而实现视频的动作识别。

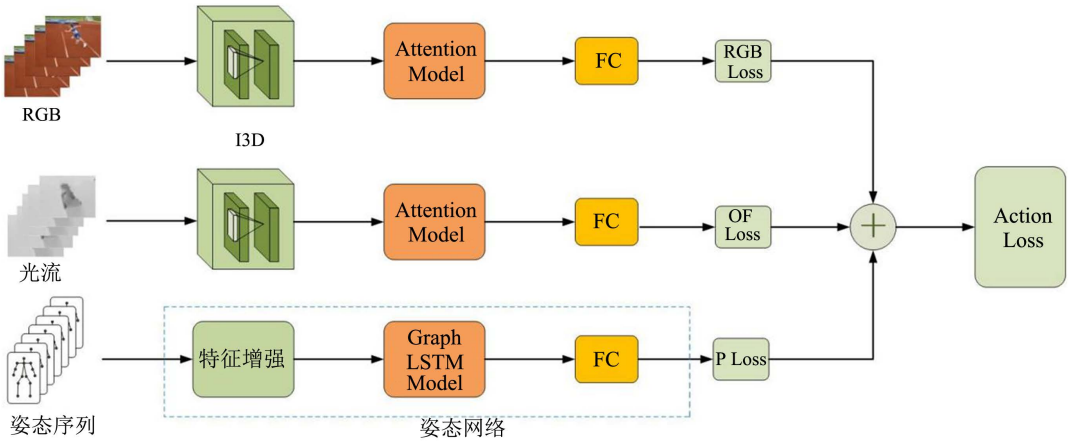
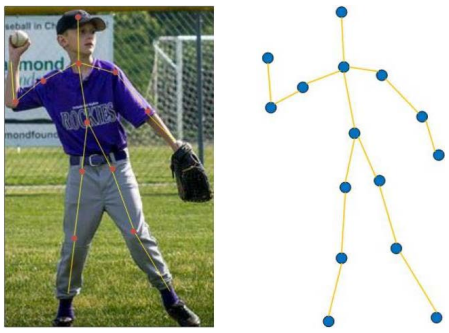


Figure 1. Overall framework of action recognition network
图 1. 行为识别网络整体架构图

3.1. 基于姿态序列的姿态图构建

人体姿态序列和关节的轨迹对场景和光照变化有很强的鲁棒性，其空间结构和时序动态变化为行为识别提供了强有力的辨别特征，在很大程度上弥补了 RGB 和光流特征的不足。通常视频由一系列的帧组成，每一帧的姿态序列是一组关节点坐标的集合，本文用姿态检测算法[6]检测视频中每一帧的人体关节的坐标位置。给定以 2D 坐标形式表示的人体关节序列，我们可以构建人体的姿态模式图。其中，以人体关节点作为图的顶点，以关节点间的生物学自然连接性作为图的边。以单帧为例，我们用 $G_t(V_t, E_t)$ 表示第 t 帧的姿态图，其中，V_t 表示人体的 N 个关节点的集合，E_t 表示生物学上关节点间自然连接的边。图 2 展示了提取的单帧姿态序列(a)和构建的以二维坐标表示的姿态模式图(b)。



(a) (b)

Figure 2. (a) Single frame pose diagram and (b) Skeleton pattern diagram
图 2. (a)单帧姿态图和(b)姿态模式图

3.2. GCN-LSTM 层

得益于强大的长时序建模能力，LSTM [27]及其变种在学习姿态序列的动态变化特征上取得了不错的效果，但是其内部的全连接特性限制了 LSTM 对姿态序列空间结构信息的表达，而这恰好是图卷积网络的优势。因此本文通过在 LSTM 中引入图卷积计算，以期实现对姿态序列的时空模式信息的学习。本文的 GCN-LSTM 单元如图 3 所示：

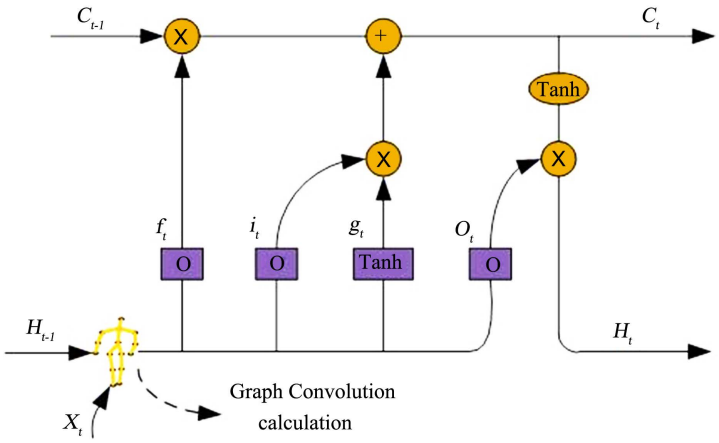


Figure 3. GCN-LSTM unit
图 3. GCN-LSTM 单元

本文用 $W_x * gX$ 表示图卷积计算。给定第 t 帧姿态图 $G_t(V_t, E_t)$，图卷积计算公式如下：

$$Y_t = D^{-\frac{1}{2}} A D^{\frac{1}{2}} X_t W \tag{1}$$

其中，X_t 表示图节点的特征，大小为 $N \times M$，N 表示图节点的数量，M 表示每个图节点的特征。A 表示整个图的空间结构的邻接矩阵，D 是对邻接矩阵进行归一化的度矩阵，W 是卷积计算的权重，Y_t 表示输入特征 X_t 在经过图卷积推理后的输出特征。

同 LSTM，我们的 GCN-LSTM 模型也有三个核心部件，分别是：输入门 i_t，输出门 o_t 和忘记门 f_t。其计算公式表示如下：

$$i_t = \sigma\left(W_{xi} * gX_t + W_{hi} * gH_{t-1} + b_i\right) \tag{2}$$

$$f_t = \sigma\left(W_{xf} * gX_t + W_{hf} * gH_{t-1} + b_f\right) \tag{3}$$

$$o_t = \sigma\left(W_{xo} * gX_t + W_{ho} * gH_{t-1} + b_o\right) \tag{4}$$

$$g_t = \tanh\left(W_{xc} * gX_t + W_{hc} * gH_{t-1} + b_c\right) \tag{5}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{6}$$

$$H_t = O_t \odot \tanh\left(C_t\right) \tag{7}$$

其中，$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别表示 sigmoid 激活函数和 tanh 激活函数，$\odot$ 表示两矩阵的哈达玛积。g_t 表示在 t 时刻候选的记忆状态。输入 X_t 是图结构的数据，受门控机制的影响，中间隐状态 H_t 和记忆细胞 C_t 也都是图结构的数据。

综上所述，中间隐状态 H_t 包含了丰富的结构信息和时序特征。在最后一个 GCN-LSTM 层，我们把每个图节点的特征聚合，计算如下：

$$AJ_t = \sum_{i=1}^{N} H_{ti} \tag{8}$$

用最终的聚合特征 AJ_t 预测动作类别分数。

3.3. 姿态网络

对于姿态序列，本文使用全连接层和 LSTM 层将每个关节的二维坐标映射到高维特征空间进行特征增强。首先全连接层将关节的坐标映射为 256 维的向量，用 $J_t \in R^{N\times 256}$ 表示，$J_{ti} \in R^{1\times 256}$ 表示每个关节点的特征，其只包含位置信息。在图模型中，J_{ti} 位置特性是有益的学习空间结构的特征。连续两帧之间的帧差特征 F_{ti} 便于 GCN-LSTM 获取姿态序列沿时间维度的长时序依赖关系。本文将这两种特征进行串联作为增强特征来丰富节点的特征信息，以兼顾各自的优势。然而，位置特征 J_{ti} 和帧差特征 F_{ti} 的拼接存在特征向量的尺度差。因此，本文采用 LSTM 层来消除这两个特征之间的尺度差:

$$FA_{ti} = LSTM\left(concat\left(J_{ti}, \left(J_{ti} - J_{(t-1)i}\right)\right)\right) \tag{9}$$

其中 FA_{ti} 是关节 i 在 t 时刻的增强特征。不同关节之间共享线性层和 LSTM。经过 LSTM 层之后，将特征增强的序列 $\{FA_1, FA_2, FA_3, \cdots, FA_T\}$ 作为节点特征送入后面的 GCN-LSTM 层。本文堆叠了三个 GCN-LSTM 层来学习姿态序列的时空模式信息。

3.4. 注意力机制学习

受到注意力机制在 NLP 中成功应用的启发，我们在网络中引用了空间注意力机制，对 I3D 网络提取的特征根据关注区域的重要程度重新分配权重。视频中动作发生区域的特征的重要程度明显高于视频场景中其他部分特征的重要程度，利用空间注意力机制对重要区域的特征分配更高的注意力得分，提高该区域的特征显著性，以此达到抑制无关冗余信息的目的，提高网络的效率和性能。本文的注意力模型如下图 4 所示，主要部分是由 8 个卷积层构成的编码器，它能够对输入特征重新分配权重。编码器的输入特征是 I3D 网络提取的特征 $X^* = \{X_1^*, X_2^*, \cdots, X_N^*\}$，$X_i^* \in R^{H\times W\times C}$。其中 X_i^* 表示第 i 帧的特征，输入的帧数为 N。我们用*表示光流或 RGB。

解码器的作用是给出重新分配权重的特征的最终得分。其计算公式如下：

$$\beta_i^* = Conv\left(X_i^*\right) \tag{10}$$

$$W_i^* = 1\Big/\left[1+\mathrm{e}^{-\beta_i^*}\right] \quad i = 1, 2, 3, \cdots, N \tag{11}$$

其中 $Conv(\cdot)$ 是基础卷积网络，β_i^* 是由卷积网络计算得到的权重向量。W_i^* 是激活函数对卷积网络计算得到的权重向量进行归一化后得到的注意力权重系数。最后将 W_i^* 分别与对应的原始输入特征相乘，得到重新分配注意力权重的特征。用公式表达如下：

$$F^* = \sum_i^N W_i^* X_i^* \tag{12}$$

F^* 即为重新分配注意力权重的特征。

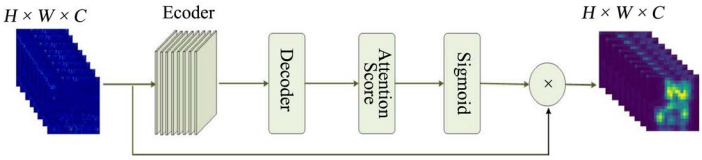


Figure 4. Attention model
图 4. 注意力模型

3.5. 损失函数

本文采用标准的交叉熵损失函数。由于模型有三个网络分支分别学习不同模态的信息，我们给每一个网络分支定义一个损失函数，以此来实现对各个分支网络的训练，经多轮训练不断迭代的产生动作分类的分类分数。对于每个网络分支，损失函数如下：

$$L_j = -\sum_{n=1}^{c}\left(y_n \log y_j\right) \tag{13}$$

其中，y_n 是动作所属类的真实标签，y_j 是网络预测的类别分数，C 表示类别数。当 $j = R$ 时表示 RGB 分支的损失函数，$j = O$ 时表示光流分支的损失函数，$j = P$ 时表示姿态分支的损失函数。

综上所述，本文模型的损失函数表示如下：

$$\text{Total Loss} = L_O + L_R + L_P \tag{14}$$

4. 实验

在本节内容中，首先介绍实验所采用的的数据集，然后说明本文的实验设置和采用的评价标准并分析论述实验结果，以验证本文方法的有效性。

4.1. 数据集

JHMDB 数据库[25]包括 21 个人类动作，收集自电影和 youtube，涉及到日常活动。完整的数据集有 928 个 clips 和 31,838 帧。数据集的训练和测试文件有 3 个，我们用训练集进行训练，测试集进行测试。数据集完整的注释了行为和姿态。包括每帧的关节位置，每个 clip 的动作标签。数据集注释丰富且轻量化，特别适合用于行为识别研究，以评估模型的性能。

4.2. 实验细节和评价标准

首先，将图片帧的分辨率调整为 112 × 112，并取连续 64 帧输入 I3D 网络提取特征，然后用注意力模块对特征进行增强，最后用三层全连接神经网络提取更高维度特征用于分类。对于姿态网络，我们采样固定长度的 T 帧姿态序列作为输入，本实验中 T = 30。在本文的 GCN-LSTM 中，每个节点的邻居集只包含与自身直接相连的节点。GCN-LSTM 层的通道设置为 512。在训练过程中，本文使用 Adam 优化器[26]对网络进行优化。初始学习率设置为 0.001，学习率的衰减率为 0.1。本文模型基于深度学习框架 Pytorch 实现。

本文采用 ACC (准确率)作为评价标准，以正确分类的个数占全部分类数的百分比来衡量模型的性能。

4.3. 实验结果及分析

本文首先消融研究在引入姿态信息后对行为识别精度的影响，如表 1 所示。接着验证注意力机制对实验结果的影响，如表 2 所示。然后将本文的模型与当前比较先进的方法作比较，如表 3 所示。

Table 1. The influence of combination of different modal information on accuracy of action recognition
表 1. 不同模态信息的组合对行为识别精度的影响

模型			ACC (%)
I3D + RGB	I3D + 光流	GCN-LSTM + 姿态序列	
√			66.4
√	√		76.3
√	√	√	83.3

Table 2. The influence of spatial attention mechanism on the recognition accuracy
表 2. 空间注意力对识别精度的影响

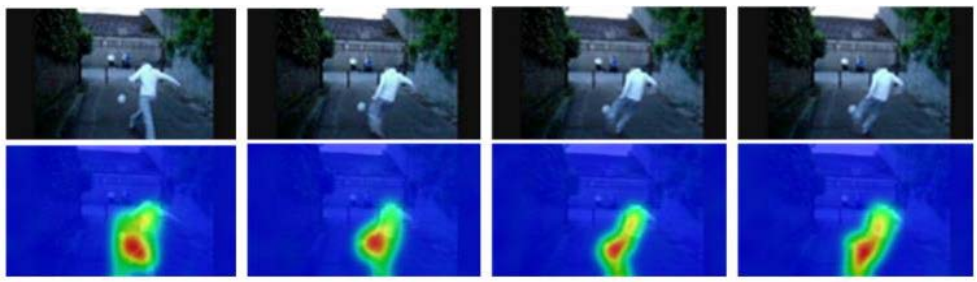
Method	ACC (%)
None	79.6
Attention	83.3

Table 3. Comparison with advanced methods
表 3. 与先进方法的比较

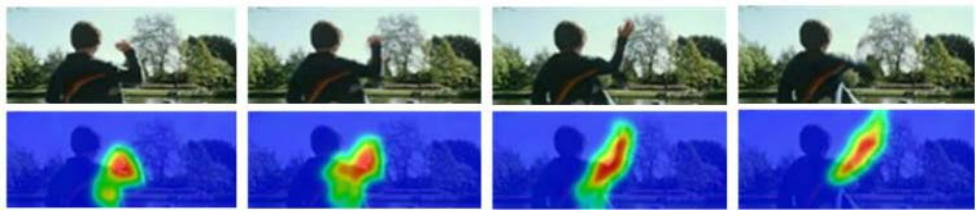
Method	JHMDB (%)
P-CNN [20]	61.1
MR Two-Stream R-CNN [21]	71.1
PA3D [22]	69.5
Chained [23]	76.1
Potion [24]	57.0
Our	83.3

由表 1 的结果可以看出：多模态信息的融合能够显著提高网络的识别能力。通常不同模态的特征包含不同行为模式的信息，但仅仅通过表观特征和运动特征去表达行为是不足的且抗干扰的能力有限。本文通过引入帧级姿态信息，并利用 GCN-LSTM 网络学习其空间结构信息和长时序依赖关系，有效的提高了识别性能。充分说明本文方法的有效性。

由表 2 可以看出，空间注意力的加入明显提高了模型的识别准确度。显而易见，本文的空间注意力机制有效的抑制了杂论的信息，增强了模型对重要区域特征的关注度，有利于对行为模式的学习。为了展示关注度的效果，本文以动作类“Kickball”和“Throw”为例，对关键帧的关注度效果进行可视化。如图 5 所示，很明显，本文的方法始终能关注动作发生的显著性区域。



(a)



(b)

Figure 5. Action “Kickball” (a) and “Throw” (b) and the attention visualization results of key frames
图 5. 动作类“Kickball”(a)和“Throw”(b)及关键帧的注意力可视化结果

表 3 的比较结果表明，相较于当前先进的方法，本文的方法有更好的表现。首先，我们的 ACC 值达到了 83.3%，明显高于其他方法。其次，相较于没有利用注意力机制的方法，本文利用空间注意力增强了模型对动作发生区域的关注度，有效的抑制了干扰。最后，本文引入姿态信息来学习行为模式，弥补了表观特征和运动特征对行为表达能力的不足，提高了网络的鲁棒性，显著提升了行为识别的性能。

5. 结束语

本文提出了一种新颖的多流网络架构来整合多种模态的特征，用于视频行为识别。首先利用 I3D 网络来学习 RGB 视频帧的表观特征和光流帧的运动特征，提取空间和时间维度的深层语义信息，并利用注意力机制对动作的发生区域重新分配权重，以此来抑制杂乱的信息，捕捉更有用的信息。在此基础上，本文引入了一个姿态网络来对人体姿态序列跨时空建模来提高网络对光照和场景变化等冗余干扰的鲁棒性，弥补表观特征和运动特征对行为表达能力的不足，提高了行为识别的准确度。在 JHMDB 数据集上的实验结果表明，本文提出的方法在识别精度上胜过当前大多数先进的方法。

参考文献

[1] Karpathy, A., Toderici, G., Shetty, S., *et al*. (2014) Large-Scale Video Classification with Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1725-1732. https://doi.org/10.1109/CVPR.2014.223

[2] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Columbus, 23-28 June 2014, 2-3.

[3] Feichtenhofer, C., Pinz, A. and Zisserman, A. (2016) Convolutional Two-Stream Network Fusion for Video Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2. https://doi.org/10.1109/CVPR.2016.213

[4] Tran, D., Bourdev, L., Fergus, R., *et al*. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[5] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2-3. https://doi.org/10.1109/CVPR.2017.502

[6] Cao, Z., Hidalgo, G., Simon, T., *et al*. (2018) OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 172-186.

[7] Laptev, I. (2005) On Space-Time Interest Points. *International Journal of Computer Vision*, **64**, 107-123. https://doi.org/10.1007/s11263-005-1838-7

[8] Scovanner, P., Ali, S. and Shah, M. (2007) A 3-Dimensional Sift Descriptor and Its Application to Action Recognition. *Proceedings of the* 15*th ACM International Conference on Multimedia*, Augsburg, 24-29 September 2007, 357-360. https://doi.org/10.1145/1291233.1291311

[9] Wang, H., Klaser, A., Schmid, C. and Liu, C. (2011) Action Recognition by Dense Trajectories. *CVPR* 2011, Colorado Springs, 20-25 June 2011, 3. https://doi.org/10.1109/CVPR.2011.5995407

[10] Wang, H. and Schmid, C. (2013) Action Recognition with Improved Trajectories. *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 3551-3558. https://doi.org/10.1109/ICCV.2013.441

[11] Wang, L., Qiao, Y. and Tang, X. (2015) Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 4305-4314. https://doi.org/10.1109/CVPR.2015.7299059

[12] Donahue, J., Anne, H.L., Guadarrama, S., *et al*. (2015) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 2625-2634. https://doi.org/10.1109/CVPR.2015.7298878

[13] Xu, K., Hu, W., Leskovec, J. and Jegelka, S. (2018) How Powerful Are Graph Neural Networks?

[14] Qi, S., Wang, W., Jia, B., Shen, J. and Zhu, S.-C. (2018) Learning Human-Object Interactions by Graph Parsing Neural Networks. *European Conference on Computer Vision*, Munich, 8-14 September 2018, 407-423. https://doi.org/10.1007/978-3-030-01240-3_25

[15] Simonovsky, M. and Komodakis, N. (2017) Dynamic Edge Conditioned Filters in Convolutional Neural Networks on Graphs. I*EEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 3. https://doi.org/10.1109/CVPR.2017.11

[16] Seo, Y., Defferrard, M., Vandergheynst, P. and Bresson, X. (2016) Structured Sequence Modeling with Graph Convolutional Recurrent Networks.

[17] Yan, S., Xiong, Y., Lin, D. and Tang, X.O. (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. 8*th AAAI Symposium on Educational Advances in Artificial Intelligence* (*EAAI*-18), New Orleans, 2-7 February 2018, 3.

[18] Li, Z.Y., Gavrilyuk, K., Gavves, E., Jain, M. and Snoek, C.G.M. (2018) VideoLSTM Convolves, Attends and Flows for Action Recognition. *Computer Vision and Image Understanding*, **166**, 41-50. https://doi.org/10.1016/j.cviu.2017.10.011

[19] Sharma, S., Kiros, R. and Salakhutdinov, R. (2016) Action Recognition Using Visual Attention. *International Conference on Learning Representations*, San Juan, 2-4 May 2016, 3.

[20] Cheron, G., Laptev, I. and Schmid, C. (2015) P-CNN: Pose-Based CNN Features for Action Recognition. *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 8. https://doi.org/10.1109/ICCV.2015.368

[21] Peng, X.J. and Schmid, C. (2016) Multi-Region TwoStream R-CNN for Action Detection. *ECCV* 2016 14*th European Conference*, Amsterdam, 11-14 October 2016, 8.

[22] Yan, A., Wang, Y., Li, Z., *et al*. (2020) PA3D: Pose-Action 3D Machine for Video Recognition. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 8. https://doi.org/10.1109/CVPR.2019.00811

[23] Zolfaghari, M., Oliveira, G.L., Sedaghat, N. and Brox, T. (2017) Chained Multi-Stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection. *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 8. https://doi.org/10.1109/ICCV.2017.316

[24] Choutas, V., Weinzaepfel, P., Revaud, J. and Schmid, C. (2018) Potion: Pose Motion Representation for Action Recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8. https://doi.org/10.1109/CVPR.2018.00734

[25] Jhuang, H., Gall, J., Zuffi, S., Schmid, C. and Black, M.J. (2013) Towards Understanding Action Recognition. *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 7. https://doi.org/10.1109/ICCV.2013.396

[26] Kingma, D.P. and Adam, J.B. (2015) A Method for Stochastic Optimization. 3*rd International Conference on Learning Representations*, *ICLR* 2015, San Diego, 7-9 May 2015, 7.

[27] Shi, X.J., *et al*. (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.

[28] Zach, C., Pock, T. and Bischof, H. (2007) A Duality Based Approach for Realtime TV-L^1 Optical Flow. *Joint Pattern Recognition Symposium*, Heidelberg, 12-14 September 2007, 214-223. https://doi.org/10.1007/978-3-540-74936-3_22

[29] Si, C., Chen, W., Wang, W., *et al*. (2019) An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019. https://doi.org/10.1109/CVPR.2019.00132