基于时空图卷积网络的视频中人物姿态分类

张懿扬[1],陈　志[1*],岳文静[2],张怡静[3]
(1. 南京邮电大学 计算机学院,江苏 南京 210023;
2. 南京邮电大学 通信与信息工程学院,江苏 南京 210023;
3. 南京邮电大学 物联网学院,江苏 南京 210023)

摘　要:为解决视频中人物姿态分类问题,提出了一种基于时空图卷积网络的改进模型。该模型首先结合人体的骨架关键点序列来构建视频中人体运动的时空特征图,将输入的视频人体骨骼关键点进行预处理,对空间节点依照人体运动规律进行子网划分,构造关节序列的时空图;继而对得到的时间特征图与空间特征图确定特征权重与卷积核,并进行级联特征融合;最后根据输入输出通道层数量搭建由图卷积网络与时序卷积网络构成的网络训练模型,基于时空特征图构型划分进行时序卷积与图卷积操作,由模型的全连接层得到分类结果。实验结果表明,上述改进模型能够准确得到视频中人物姿态的分类结果,并改善了卷积网络在训练中的特征冗余问题,有效地提高人物姿态分类的鲁棒性。

关键词:人物姿态分类;特征融合;时空图卷积网络;骨骼关键点;特征冗余
中图分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1673-629X(2021)10-0070-06
doi:10.3969/j.issn.1673-629X.2021.10.012

Human Pose Classification in Video Based on Spatial Temporal Graph Convolutional Networks

ZHANG Yi-yang[1],CHEN Zhi[1*],YUE Wen-jing[2],ZHANG Yi-jing[3]
(1. School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China;
2. School of Telecommunications and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210023,China;
3. School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:In order to solve the classification problem of human pose in videos,an improved model based on spatial temporal graph convolution network is proposed. In this model,firstly the human skeleton key point sequences are combined to construct a spatial-temporal feature map of human motion in the video. Openpose is used to preprocess the input skeleton key point data in the video,and subnets are divided from spatial construction according to the rule of human motion to obtain a spatial-temporal feature map of the joint sequence. Then feature weights and convolution kernel are determined for the obtained spatial-temporal feature maps,and feature fusion is carried out in cascade. Finally,according to the number of input and output channel layers,a training model composed of the graph convolution network and the temporal convolutional network is built. The temporal convolution and the graph convolution are performed based on the configuration division of the spatial-temporal characteristic graph,and the classification results can be obtained from the full connection layer of the model. The experiment shows that the improved model can accurately obtain the classification results of the characters in the video,and improve the feature redundancy of the convolutional network in the training,thus effectively improving the robustness of the classification of characters.

Key words:human pose classification;feature fusion;spatial temporal graph convolutional networks;skeletal key point;feature redundancy

收稿日期:2020-07-22　　**修回日期**:2020-11-23
基金项目:江苏省重点研发计划(社会发展)项目(BE2016778,BE2019739);南京邮电大学科研项目(NY217054);南京邮电大学大学生创新训练计划项目(XZD2019073)
作者简介:张懿扬(1998-),女,研究方向为软件工程;通信作者:陈　志(1978-),男,教授,CCF 会员(14587M),研究方向为移动物联网、无线传感器网络、数据挖掘。

0　引　言

在进行人物姿态分类时,一般认为人的行为具有多种模式,在群体人物行为中表现为三个或者更多相互作用、相互影响、有共同目标的人物组成的群体的相对运动现象[1],具体到竞技比赛中表现为运动者做出一些具有典型特点的竞技运动行为。通过对运动特征的分析与处理来提出可行性建议,可以提高运动质量,优化人们生活品质,提升体育竞技水平[2]。当前已开始使用深度学习等算法进行研究实验,形成了基于人工特征提取[3]的传统方法和基于深度学习[4]的方法。

在 Microsoft Kinect、OpenPose 等人体姿态检测系统中,人体关键点的运动轨迹为动作的描述提供了较好的表征,基于骨架关键点的模型通常能传达出重要的特征信息。以连续的视频帧中检测到的人体骨架关键点序列为输入,能够输出在视频中发生的人物动作类别。在早期使用骨骼进行人物动作识别时,仅仅利用单个时间步长的关键点坐标形成的特征向量,对其进行时间分析,并没有显式地利用关键点之间的空间联系。

在近期的研究中,利用关键点间自然联系的方法已经被开发出来[5],其效果较早期方法有着长足的进步,证明了连接的重要性。目前多数方法仍需要用手工制作的规则及部件来分析空间模式[6],这导致应用程序的设计模型很难泛化使用。

Yan S 考虑到骨架的结构特征,提出了新的动态骨架模型:ST-GCN[7](时空图卷积网络)。因为其以图结构而非网络形式呈现,给卷积网络模型的使用造成困难。近来发现可以对任意图结构进行卷积的 GCN(图卷积网络)有了快速的发展,不同于传统的基于图像的 CNN(卷积神经网络),GCN 可以基于任意拓扑结构进行卷积,因此可以基于人体姿态估计构造拓扑的人体结构,对人体结构进行运动学分析,最后以时序的人体结构进行 CNN 卷积,基于卷积的结果进行人体运动姿态分类。

尽管人体运动姿态分类任务有了新方向,在进行卷积网络模型训练时常常会出现特征冗余的问题。在卷积网络训练的过程中,关键点检测是检测人体的不同部分,并不是人体各个部分的特征都集中在最后一层特征图上,不同部分的特征可能会分布到不同尺度的特征图上,如果只是通过最后一层的特征图来进行关键点检测,会得到比较差的结果。Feichtenhofer 等[8]在分类器级融合[9]的基础上提出了时间与空间特征融合算法,通过在 Softmax 层和卷积层之后的 ReLU 层进行操作以实现特征级的融合,有效避免特征冗余的问题,增强模型的鲁棒性。

1　基于 ST-GCN 的模型分析

考虑到要处理视频时需要分别进行空间部分与时序部分的网络训练,引入 ST-GCN(spatial temporal graph convolutional networks)模型,即,该模型是结合图卷积网络和时间卷积网络进行的基于骨骼点的动作识别分类模型。

首先,为得到骨骼的数据,先使用 OpenPose 对视频进行预处理。该步骤将人体关节连接成骨骼,从而进行姿态估计。ST-GCN 模型只需要关注其输出,由一系列的输入帧的身体关节序列构造时空图,时空图中将人体的关节作为图节点,将人体结构和时间上的连通性对应的两类边组成图的边。因此,ST-GCN 将图节点的联合坐标向量作为输入,再对数据进行多层时空图卷积操作,从而形成更高级别的特征图。最后由 SoftMax 分类器分类到相应的动作类别。

在描述特征时需要表达多层骨骼序列从而尽可能保留原有的信息,ST-GCN 使用时空卷积图来描述,能够很好地做到这一点。该模型使用每一帧每一个人体骨骼的坐标表示骨架关键点序列,基于此构建一个时空图,其中人体的关节关键点为图的节点 V,以骨架自然连接方式构建空间图,得到空间边集 E_{sp};身体结构的连通性和时间上的连通性为图的时序边集 E_t。由以上节点集与空间边集可构成训练时所需要的时空特征图。

在该模型中使用的卷积公式为:

$$\mathrm{aggre}(x) = D^{-1}AX \tag{1}$$

其中,$\hat{A} = D^{-1}A$ 被看作为卷积核。

当进行卷积操作时,因为动作识别的特点使用了图划分而非单一的卷积核进行操作,将 $\hat{A}$ 分解为 $\hat{A}_1$,$\hat{A}_2$ 与 $\hat{A}_3$,其中 $\hat{A}$ 表示的边具有以下特点:两个节点之间有一条双向边;节点自身有一个自环。结合运动规律分析,该模型将卷积核分为三个子图,分别用来表示向心运动、离心运动与静止动作特征。对于一个根节点,具体划分方式为:

第一部分连接根节点本身,表示静止的特征;第二部分连接比节点本身更靠近骨架重心的相邻节点集,表示向心运动的特征;第三部分连接比节点本身更远离骨架重心的相邻节点集,表示离心运动的特征。

由此,一个卷积核将被划分为三个,其中每个卷积结果都代表了不同尺度的动作运动特征,将它们进行加权平均便能得到卷积的结果。

在完成以上步骤后,可以得到带有 k 个卷积核的图卷积表达式:

$$\sum_k \sum_v (XW)_{\mathrm{nkctv}} \hat{A}_{\mathrm{kvw}} = X'_{\mathrm{nctw}} \tag{2}$$

式中，n 表示视频中的人数，k 表示卷积核数量，使用上述子图划分方法后一般值为 3，c 表示关节特征数，t 表示关键帧数，v、w 表示 OpenPose 预处理后的输出关节数。对 v 求和代表节点的加权平均，对 k 求和代表不同卷积核的加权平均。

在通过图卷积网络得到空间上的节点特征后，还需要学习得到时间中节点的变化特征。在该模型中使用 TCN，即时间卷积网络，通过传统卷积层的时间卷积操作来完成这项任务。

与此同时，考虑到在运动过程中，不同部分的躯干变化特征的重要性各不相同，因此在每个 ST-GCN 单元中都引入了注意力模型[10]，用于在最后加权计算时保证每个单元有自己合适的权重，提高模型训练的准确性。最终利用标准的 Softmax 分类器将合并所有时空特征后的特征图分类到相应的类别当中，并用梯度下降 SGD 进行结果误差优化。

2 基于时空卷积网络的视频人物姿态分类

时空卷积网络 ST-GCN 具有良好的人体运动动作分类能力，但该模型仍存在不足之处。经过特征提取后得到的特征数据也存在着一些问题，例如，一单元内的方差较小，整体数据各个单元之间却有着较大数值的方差，而因为对图像部分特性的变化和其他特性的变化相比，同一种特征的敏感性不相同，因此当两类图像差异在某种特征敏感特性上的差异不大时，基于单一特征训练的分类器将无法得出正确的分类。此外，其他问题也会导致特征数据质量下降，例如复杂的背景噪声等，既增加了分类器训练难度，也降低了其准确性。

特征融合的方法可以解决这种问题，这种方法同时提取多种特征进行分类器训练，实现特征的互补。Simonyan 提出了一种使用双流架构的网络模型[9]，该模型属于分类器级的融合，通过对建立的空间流和时间流卷积神经网络分别进行独立训练，最终由 Softmax 分类器将上述两个网络融合。而 Feichtenhofer 在上述模型基础上做了网络融合方面的改进[8]，提出了空间特征融合和时间特征融合算法，可以同时在分类器级和 ReLU 层进行融合操作，实现特征级的融合。

ST-GCN 模型目前仅适用于分类器级的融合，因此对该模型进行如下改进：

针对该模型的时空特征图进行特征的融合，融合函数定义为[8]：

$$f: x_t^a + x_t^b \rightarrow y_t \tag{3}$$

式中，x_t^a 和 x_t^b 表示在 t 时间点视频帧经过卷积运算输出的时空特征图，y_t 表示融合时空特征图，$x_t^a, x_t^b, y_t \in \mathbb{R}^{H\times W\times D}$，$H$、$W$ 与 D 分别表示特征图的长度、宽度与通道数。融合函数通过这种方法将卷积层输出的 2 个特征图进行融合，从而将两个卷积神经网络连接起来，连接点称为融合点。

融合函数包括加性融合函数、级联融合函数、卷积融合函数等，为了尽可能保留原本特征图的结果，此处使用级联融合函数 $y^{\text{cat}} = f^{\text{cat}}(x^a, x^b)$，即：

$$y_{i,j,2d}^{\text{cat}} = x_{i,j,d}^a, y_{i,j,2d-1}^{\text{cat}} = x_{i,j,d}^b \tag{4}$$

式中，$y \in \mathbb{R}^{H\times W\times 2D}$，$i \in [1, H]$，$j \in [1, W]$，$d \in [1, D]$。它能够保留两个特征图的结果，保证特征描述的准确性，而融合后特征图的通道数量会变为原始特征图的两倍。实现特征级的融合，增强分类器结果的鲁棒性[11]。

3 实验与结果分析

3.1 实验准备

实验的数据集来源于 hmdb51 数据集[12]，该数据集是当前识别动作研究领域最为重要的几个数据集之一，多数来源于电影，还有一部分来自公共数据库以及 YouTube 等网络视频库，包含 51 类动作，共有 6 849 个视频，每个动作每类至少包含有 101 段样本，分辨率为 320 * 240。本次实验采用该数据集中的肢体动作部分，包括一般身体动作（general body movements）、与对象交互动作（body movements with object interaction）和人体动作（body movements for human interaction）三大类，实验数据集参数如表 1 所示。

表 1 实验数据集参数

参数名称	参数含义	数值
C	关节点特征值个数	3
T	视频帧数	300
V	骨架节点个数	18
M	人数	2
N	视频个数	Batch_size

3.2 实验过程

为证明基于 ST-GCN 的视频人物运动分类改进模型的有效性以及更好的鲁棒性，现对某一视频进行运动人物检测与跟踪。该视频分辨率为 320 * 240，帧

率为 30 F/s。

前 10 帧中每隔一帧取一次图像输出的运动人物检测与跟踪结果,如图 1 至图 5 所示,分类结果为上述图中人物正进行射门(足球运动)。

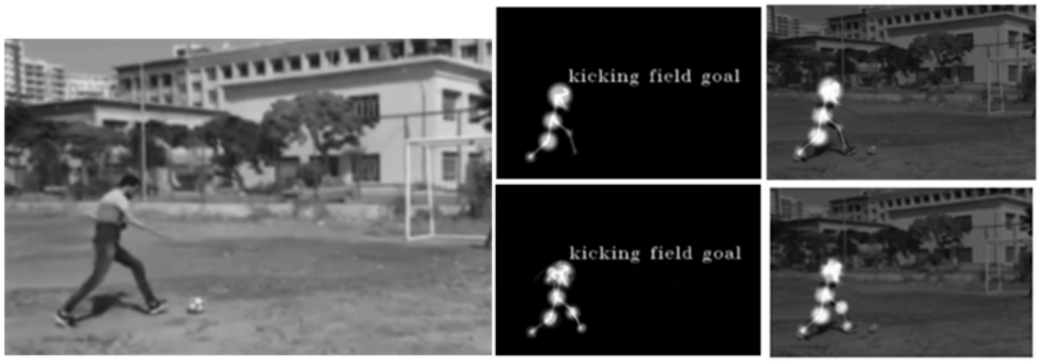

图 1　第 1 帧图像(左)与运动姿态分类结果

(中上为原 ST-GCN 模型输出的骨骼图,中下为改进后模型骨骼图;右上为原 ST-GCN 模型输出的 RGB 渲染图,右下为改进后模型渲染图)

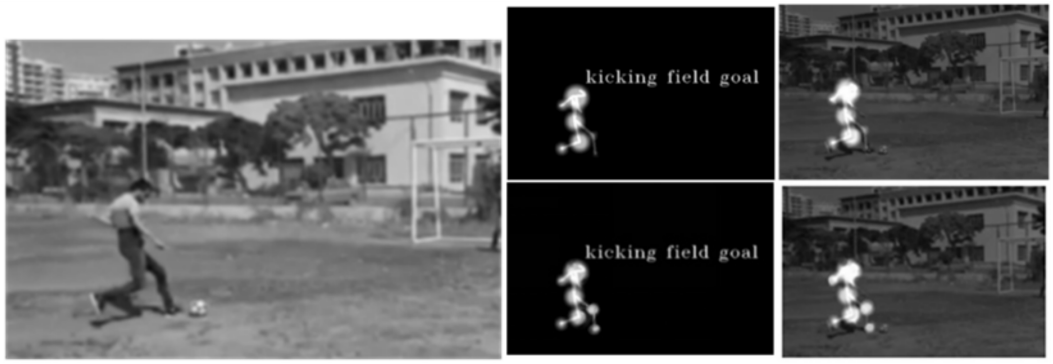

图 2　第 3 帧图像(左)与运动姿态分类结果

(中上为原 ST-GCN 模型输出的骨骼图,中下为改进后模型骨骼图;右上为原 ST-GCN 模型输出的 RGB 渲染图,右下为改进后模型渲染图)

图 3　第 5 帧图像(左)与运动姿态分类结果

(中上为原 ST-GCN 模型输出的骨骼图,中下为改进后模型骨骼图;右上为原 ST-GCN 模型输出的 RGB 渲染图,右下为改进后模型渲染图)

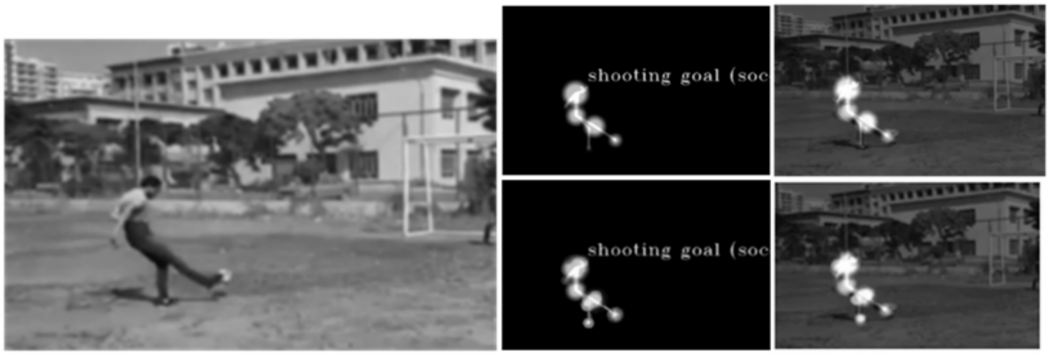

图 4　第 7 帧图像(左)与运动姿态分类结果

(中上为原 ST-GCN 模型输出的骨骼图,中下为改进后模型骨骼图;右上为原 ST-GCN 模型输出的 RGB 渲染图,右下为改进后模型渲染图)

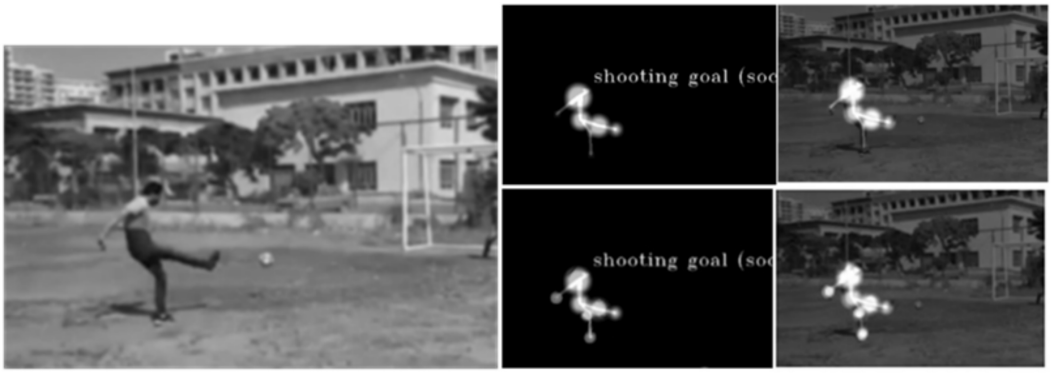

图 5 第 9 帧图像(左)与运动姿态分类结果

(中上为原 ST-GCN 模型输出的骨骼图,中下为改进后模型骨骼图;右上为原 ST-GCN 模型输出的 RGB 渲染图,右下为改进后模型渲染图)

3.3 结果分析

实验采用 HMDB51 数据集进行测试,将结果根据动作分类进行每一类的对比后,再将改进后的模型与双流法 Two-stream 模型、深度监督网络 Deeply-Supervised Nets 模型[13]、残差网络 ResNet-152[14]模型、TVNet[15]模型进行了对比,结果如表 2 和表 3 所示。

表 2 模型改进前后不同动作的准确性比较 %

动作分类	改进模型	原模型
Pullup	70.45	70.16
KickBall	70.77	69.35
BendingBack	69.61	68.89
ClimbingTree	69.53	68.92
MilkingCow	70.29	69.35
HittingBaseball	70.23	69.75
ClianingWindows	70.53	68.47
平均值	70.2	69.27
方差	0.19	0.28

表 3 改进模型与其他模型的准确度比较 %

模型	HMDB51
TVNet+IDT[16]	72.6
ResNet-152[14]	63.8
Two-Stream Fusion[17](VGG-16)	65.4
Deeply-Supervised CNN Model + optical flflow[18]	73.9
Transformation[19](VGG-16)	62.0
原模型	69.3
改进模型	70.2

图 1 至图 5 展示了视频中人物运动姿态分类模型的姿态分类结果,结果显示原 ST-GCN 与改进后的模型均能够正确地分类视频中的人物动作,但是相比于原模型,改进后的模型对于每个关节节点的特征进行卷积时的权重分配更符合人体运动规律。从表 2 和表 3 中能够看出,改进模型改善了原模型特征类内方差较小而类间方差较大的问题,具有更好的鲁棒性。

4 结束语

文中提出了基于时空图卷积网络的视频中人物姿态分类改进模型,该模型首先将输入的视频人体骨骼关键点进行处理,构造关节序列的时空图,继而将得到的时间特征图与空间特征图分别划分并进行级联特征融合,最后基于空间构型划分进行时序卷积与图卷积操作,得到分类结果。该模型能够准确得到视频中人物姿态的分类结果,解决了卷积网络在训练中的特征冗余问题,有效地提高人物姿态分类的鲁棒性。

在后续的研究中,可以考虑对视频中的一个群体进行动作分类,此外该算法是针对完整目标检测后的特征图分析与训练结果的分类,对于画面中有不完整人物出现时的目标检测仍有较大的发展空间。

参考文献:

[1] YAN R, TANG J, SHU X, et al. Participation-contributed temporal dynamic model for group activity recognition [C]//Proceedings of the 2018 ACM international conference on multimedia. New York: ACM, 2018: 1292-1300.

[2] 刘昊扬. 基于人工智能的运动教练系统分析与展望[J]. 北京体育大学学报, 2018, 41(4): 55-60.

[3] 温长吉. 行为识别中特征提取和描述相关问题研究[D]. 长春: 吉林大学, 2017.

[4] 樊 恒, 徐 俊, 邓 勇, 等. 基于深度学习的人体行为识别[J]. 武汉大学学报: 信息科学版, 2016, 41(4): 492-497.

[5] WANG J, LIU Z, WU Y, et al. Mining actionlet ensemble for action recognition with depth cameras[C]//2012 IEEE conference on computer vision and pattern recognition. Providence: IEEE, 2012: 1290-1297.

[6] 罗会兰, 王婵娟, 卢 飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169-180.

[7] YAN S, XIONG Y, LIN D. Spatial temporal graph convolu-

tional networks for skeleton-based action recognition[C]// Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18). New Orleans: AAAI, 2018: 7444-7452.

[8] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR2016). Las Vegas: IEEE, 2016: 1933-1941.

[9] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 27th international conference on neural information processing systems advances in neural information processing systems (NIPS´14). Cambridge, MA, USA: MIT Press, 2014: 568-576.

[10] 孟乐乐. 融合时空网络与注意力机制的人体行为识别研究[D]. 北京:北京交通大学,2018.

[11] 刘渭滨,邹智元,邢薇薇. 模式分类中的特征融合方法[J]. 北京邮电大学学报,2017,40(4):1-8.

[12] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]// Proceedings of the 2011 international conference on computer vision. Barcelona: IEEE, 2011: 2556-2563.

[13] LEE C Y, XIE S, GALLAGHER P, et al. Deeply-supervised nets[C]/Proceedings of the 18th international conference on artificial intelligence and statistics. San Diego: AISTATS, 2015: 562-570.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR2016). Las Vegas: IEEE, 2016: 770-778.

[15] FAN L, HUANG W, GAN C, et al. End-to-end learning of motion representation for video understanding[C]//Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 6016-6025.

[16] CHO S, FOROOSH H. A temporal sequence learning for action recognition and prediction[C]//Proceedings of the 2018 IEEE winter conference on applications of computer vision (WACV). Lake Tahoe: IEEE, 2018: 352-361.

[17] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal multiplier networks for video action recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 4768-4777.

[18] LI Y, LI K, WANG X. Deeply-supervised CNN model for action recognition with trainable feature aggregation[C]// Proceedings of the international joint conference on artificial intelligence. Stockholm, Sweden: IJCAI, 2018: 807-813.

[19] WANG X, FARHADI A, GUPTA A. Actions ~ transformations[C]//Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR2016). Las Vegas: IEEE, 2016: 2658-2667.