Computer Science and Application 计算机科学与应用, 2022, 12(8), 1877-1886
Published Online August 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.128188

Hans汉斯

基于局部特征移位网络的手部骨架动作识别

田文浩，陈俊洪，钟经谋，刘文印

广东工业大学计算机学院，广东 广州

收稿日期：2022年7月1日；录用日期：2022年7月30日；发布日期：2022年8月5日

摘　要

由于视觉的不稳定性和环境的复杂性，基于第一人称视角的动作行为难以得到准确的识别。本文提出了一种基于局部特征移位网络的动作识别网络框架，具体来说，该框架首先建立手部骨架的无向时空图拓扑结构，并使用自适应图卷积网络提取手部骨架拓扑图的关节特征和连接信息；其次，为了得到全局空间信息，本文使用ResNet152网络提取RGB特征。在获得手部骨架特征与RGB图像特征后，我们将其分别输入到提出的局部特征移位卷积网络，该网络通过样本间的互相学习为模型带来更好的泛化性。通过在FPHA数据集上进行的实验表明，该框架在动作识别上的精确度证明了该模型能够有效地应对视频背景干扰，并具有较强的鲁棒性。

关键词

动作识别，图卷积，特征移位，手部骨架

Hand Skeleton Action Recognition Based on Local Feature Shift Network

Wenhao Tian, Junhong Chen, Jingmou Zhong, Wenyin Liu

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Jul. 1st, 2022; accepted: Jul. 30th, 2022; published: Aug. 5th, 2022

Abstract

Due to the visual instability and the complexity of the environment, first-person action is difficult to recognize accurately. In this paper, we propose an action recognition network framework based on local feature shift networks. Specifically, the framework first builds the undirected spatio-temporal graph topology of the hand skeleton, and uses adaptive graph convolution network to extract joint features and connection information of the hand skeleton topology; after that, in order to obtain

文章引用: 田文浩, 陈俊洪, 钟经谋, 刘文印. 基于局部特征移位网络的手部骨架动作识别[J]. 计算机科学与应用, 2022, 12(8): 1877-1886. DOI: 10.12677/csa.2022.128188

global spatial information, we use ResNet152 network to extract RGB features. Getting the hand skeleton features and the RGB image features, we input them into the proposed local feature shift convolutional network respectively where through the mutual learning between samples, the model could receive better generalization. Experiments on FPHA data set show that the proposed framework is accurate in motion recognition, which proves that the model can effectively deal with video background interference and has strong robustness.

Keywords

Action Recognition, Graph Convolution, Feature Shift, Hand Skeleton

1. 引言

近年来，随着人们生活的多样化，应用穿戴式相机拍摄被广泛地应用在日常生活。目前，基于穿戴式相机的第一人称视频动作识别已成为计算机视觉研究的热点问题。与第三人称视角的相机相比，第一人称视角的相机可以捕捉到穿戴者的操作画面中心，无遮挡地拍摄到被操纵的物体和工具，并且能够跟随相机穿戴者的移动而移动视角。因此，基于第一人称视角的视频拍摄更贴近人类的视觉观察，为智能机器人的发展提供了帮助。

由于可穿戴相机的移动带来的视频画面晃动以及复杂背景的干扰，如何准确地进行动作的识别仍然是一个难题。对此，相关研究者提出许多识别动作的方法，如 Singh 等人[1]提出基于特征轨迹的第一人称视角的动作表示方法。Kwon 等人[2]引入了基于 RGB 和光流特征的双流卷积神经网络，通过长期时间融合池化操作提高动作识别精度。Lu 等人[3]使用一个基于外观和运动状态的双流深度神经网络来识别第一人称视角的动作。Tang 等人[4]利用数据流信息不同的特性，设计了多流深度神经网络 MDNN，该网络保留每个数据形态的独特属性，同时探讨共享的特征，提高了识别的准确性。然而，这些方法大部分利用表面外观信息和运动流信息，对于相机移动带来的画面影响处理效果不佳，并且处理的速度也比较慢。

近年来，基于人体骨架模型的动作识别方法越来越受到广泛关注和实际应用[5] [6]。Li 等人[7]提出应用时空图卷积网络 ST-GCN 对人体骨架进行建模，通过自动学习数据的空间和时间序列的特征，克服了以往依赖人工定义骨架遍历规则方法的不足。但是 ST-GCN 建模的图结构在所有输入样本和网络层上都是固定的，Shi 等人[8]提出了改进的自适应图卷积网络 2s-AGCN，图的拓扑结构可以单独学习与优化，以适应不同的数据样本和网络层，增加了构建图的灵活性。Su 等人[9]提出一种新的基于编码器－解码器的循环神经网络用于无监督的骨架动作识别，可以在不提供标签的条件下通过聚类实现骨架序列与动作的关联。虽然以上方法利用骨架模型实现动作的精准识别，但是这些方法是基于人体全身结构的行为动作识别，对于局部的手部信息并没有充分利用。

在日常活动中，人类通常通过双手来抓取物体及操作工具，手部信息携带有重要的动作识别信息。因此，本文提出利用手部的拓扑结构进行动作识别。在动作分析中，为了克服以骨架信息作为线索的方法的不足，本文提出了一种结合骨架数据流和 RGB 图像特征的方式来进行动作识别。具体来说，我们首先将手的物理骨架结构通过拓扑图表示，然后利用自适应图卷积网络对手骨架序列进行动作特征提取，

最后输入到提出的移位卷积中进行处理。在 RGB 帧数据上，我们采用 ResNet152 网络提取 RGB 中的空间特征并利用移位卷积进行特征的进一步分析。最后我们将两个网络分支的结果进行相加，并输出最终的动作结果。

本文的主要贡献如下：

1) 本文基于手部骨架构建自适应的图拓扑结构，该结构可以根据不同的网络层和样本来优化参数，提高模型的灵活性。

2) 本文提出了特征移位卷积模块，使得不同样本间特征信息能够互相共享，提高网络识别精度；

3) 通过在大规模第一人称手－动作视频 FPHA 数据集上进行实验，表明本文所提出的网络框架可以取得较大的性能提升。

2. 相关工作

2.1. 基于 RGB 和光流信息的动作识别

在第一人称视频中，由于看不到人体本身的运动，所以只能通过手部信息来对动作和交互活动进行识别。Singh 等人[10]提出一个三流框架网络，在第一流中使用第一人称线索的自我卷积神经网络(Ego ConvNet)，同时用空间流和时间流特征的双流框架扩展模型网络。Shuichi 等人[11]提取分割后的手部掩模分别输入到二维的卷积网络和三维的卷积网络，然后利用两个网络输出的不同维度特征进行融合，证明了多流方法比单流的方法具有更好的结果。Tang 等人[12]提出在端到端的多流深度神经网络中使用手部信息作为辅助流可以更好地提升动作识别效果。然而，基于图像序列的动作识别往往需要从图像中识别出关键的特征信息，过滤大量的背景信息，容易受到噪声的干扰，对于机器的性能要求较高，计算量大。

2.2. 基于骨架的动作识别

随着图卷积网络的提出与应用，人们尝试将图卷积应用在骨架序列上，并取得了不错的效果。Liu 等人[13]在图卷积的基础上提出了多尺度聚合方案，利用统一时空图卷积网络 G3D 直接从骨架图序列建模时空依赖关系，时空特征的多尺度聚合提高了模型的性能。Shi 等人[14]提出一种新的多流自适应图卷积神经网络，该网络既可以端到端输入数据进行统一地学习，又可以单独学习，增加了图构建模型的灵活性。Xia 等人[15]设计了多层次混合时序卷积模块，通过不同尺度卷积核的组合，提供了灵活的时序图。Zhang 等人[16]提出基于编码器－解码器的神经网络用于无监督的骨骼的动作识别系统。Peng 等人[17]通过去除不同网络层中的冗余拓扑图，提供更好的消息聚合。

动态的骨架序列通常可以传达非常重要的信息，例如 Cai 等人[18]利用骨骼关节周围的视觉信息，有效捕获有用的局部微小的身体运动线索；Xie 等人[19]提出跨通道图卷积网络，使用特征融合机制从不同特征通道更新根结点的特征；动作识别与三维姿态估计有着高度的相关性，在基于骨架的动作识别[20][21]中，关节的位置信息用于识别动作类别。Das 等人在[22]将三维的人体姿态和 RGB 线索投射到一个共同的语义空间中，并通过动作识别框架提取这两种模式的信息学习时空特征。Yang 等人[23]提出了一种新的协作学习网络用于动作识别和三维手部姿态估计，该网络利用关节感知特征，采用一种新的多阶多流特征分析方法，能够有效地从视频中间特征图中学习姿态和多阶运动信息。

然而，上述方法大部分基于人体骨架进行建模和动作识别，并没有充分利用其它的图像特征，在动作感知的过程中也没有考虑手部不同部位关节之间的联系。本文利用手部关节建模骨架图，结合图像特征完成手部动作识别，下面将详细介绍方法细节。

3. 本文方法

如图 1 所示，本文提出的框架图分为三个模块：手部骨架图构建模块，特征处理模块和特征移位卷积模块。其中，手部骨架图构建模块主要对输入视频提取手部骨骼关节点信息；特征提取模块分别对手部骨架序列和 RGB 图像进行特征提取，对于 RGB 图像特征我们使用 ResNet152 网络进行特征提取，而骨架序列特征使用自适应图卷积网络进行处理；在提取到手部骨架特征与 RGB 图像特征之后，我们分别将其输入到特征移位卷积模块并分别得到动作特征分数，然后再将两个特征分数相加，最后获得预测的动作分类标签。接下来将具体介绍框架的细节。

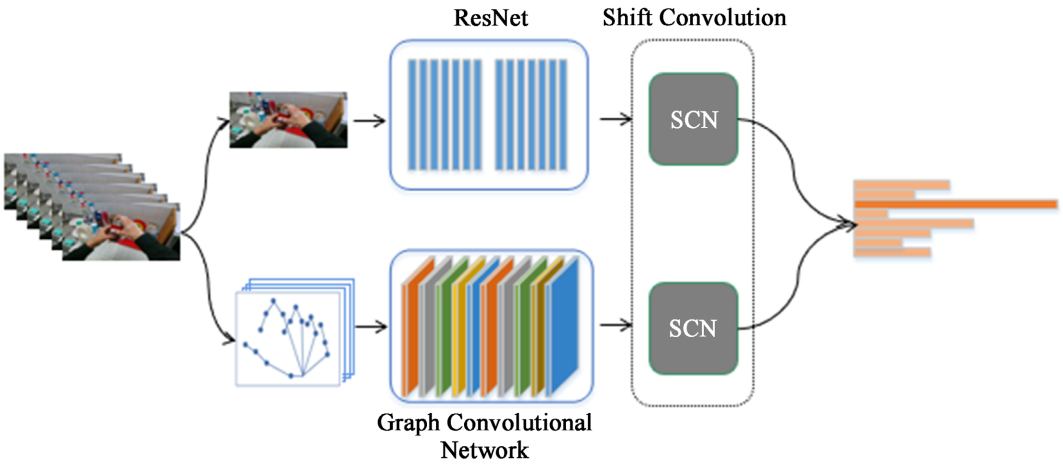

Figure 1. Network architecture
图 1. 网络框架图

3.1. 骨架序列特征处理算法

3.1.1. 手部骨架时空图

对于输入的每一帧图像，都有采集对应的手部骨架信息，包括关节序列的三维坐标位置，同一帧维度内的手部骨架由 21 个关节点组成，其中，每根手指由 4 个关节点定位。由于操作动作包含时间和空间上的连续变化，所以我们运用时空图结构来模拟这些关节沿空间和时间维度上的结构化信息。具体来说，根据手部的物理结构，我们设计了手骨架时空图，可视化效果如图 2 所示。在空间维度上，同一维度的蓝色顶点表示为每只手的关节，蓝色边表示为在空间物理结构上两个关节点的自然连接。在时间维度上，

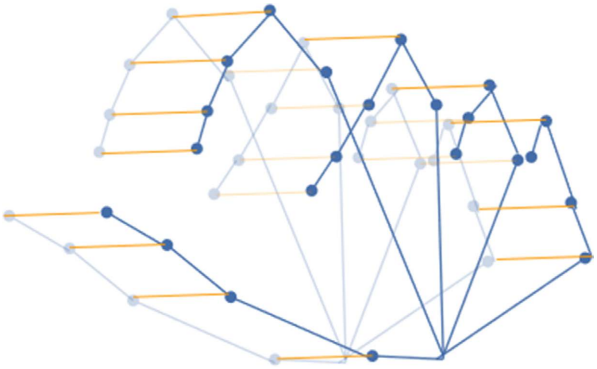

Figure 2. Spatial-temporal graph of hand skeleton
图 2. 手部骨架时空图

使用黄色边表示两个相邻帧之间的对应关节顶点的连接。对于每个关节顶点，设置坐标向量作为顶点的特征属性，可以表示为 $\mu_i=(x_i,y_i,z_i)$。

更进一步地，我们根据物理连接结构，将一个具有 N 个关节和 T 帧的骨架序列表示为一个无向时空图 $G=(V,E)$。其中，V 表示顶点集，包含了 N 个关节的特征属性，每一帧骨架数据都是手关节的三维坐标构成的向量序列，其表示为：

$$V=\{\mu_{ti}\mid t=1,\cdots,T,i=1,\cdots,N\} \tag{1}$$

μ_{ti} 表示骨架序列中第 t 帧第 i 个关节点数据；E 是描述两个节点间的连接边的关系。其中，在空间维度上同一帧内相邻顶点的连接边表示为：

$$E_m=\sum_{i,j}^{N}\mu_{ti}\mu_{tj} \tag{2}$$

时间维度上同一节点在相邻帧之间的连接边表示为：

$$E_T=\sum_{t}^{T}\mu_{ti}\mu_{(t-1)i} \tag{3}$$

3.1.2. 自适应图卷积

图卷积网络模型包括两个部分：空间图卷积和时间图卷积。在空间图卷积方面，我们根据物理结构连接关系建立关节结点邻接矩阵 $A\in\{0,1\}^{N\times N}$，其中 N 表示关节点数。为了确定图卷积的空间位置，我们将结点及其邻域构成的集合分为三个分区 Z{向心区，根结点，离心区}。利用时空图卷积可以有效地提取时空信息，但是却存在以下两个不足：1) 利用物理连接结构构造的图虽然能表达结点与邻居结点的依赖关系，但是忽略了在动作执行过程中非物理连接的手关节之间的相互影响。例如，当识别“拍手”动作时，两只手之间关节的互动比单只手内的关节更密切；“抓取”一个物体时，拇指和中指往往距离很近，它们关节之间的关系比其它关节更重要。而时空图卷积网络 ST-GCN 使用预定义的图数据结构进行计算，很难捕捉到这种关系。2) 在动作识别任务中，使用的多层图神经网络卷积层和样本往往是不同的，而 ST-GCN 只能处理固定图大小，无法进行自我调整。

为了解决上述问题，本文使用一种自适应卷积网络模块，该卷积模块对图的邻接矩阵进行了拓展，可以使图的拓扑结构随着不同的样本和网络层进行自优化，大大提高了模型的灵活性。其公式如下：

$$F_{out}=\sum_{k}^{K_v}W_kF_{in}\left(A_k+B_k+C_k\right) \tag{4}$$

其中 A_k 表示根据手的自然连接建立的拓扑空间结构，用于描述相邻关节结点的连接，与公式 4 中提到的 $N\times N$ 矩阵 A 相同，其在训练过程中是固定的。B_k 也是 $N\times N$ 的邻接矩阵，可以在训练过程中随着网络的反向传播而调整和优化，B_k 的加入使得在 A_k 中未有关联的节点建立新的新连接，让网络可以针对特定的数据进行图结构的学习，更具有灵活性。C_k 针对每个样本学习一个唯一的图，确定样本中每个顶点之间是否存在连接以及连接的强度。

时间域上的卷积与文献[7]相同，使用规则的卷积操作。具体来说，在 $T\times N$ 的特征向量图上，选取采样长度为 $K_t\times 1$，对同一顶点特征在 $K_t\times 1$ 时间帧范围内卷积，表示为：

$$H_{out}=\sum\nolimits_{K_t}H_{in}W_T \tag{5}$$

3.2. RGB 图像特征处理算法

本文采用 ResNet152 网络[24]对 RGB 图像进行特征的提取，具体来说，我们挑选具有代表性的图像

作为网络的输入，并将该图像大小裁剪为 224 × 224 × 3，紧接着将其输入到 ResNet152 网络中进行训练。其中，如图 3 所示，ResNet152 由基于 Bottleneck 的卷积块搭建而成。每个 Bottleneck 包含三个卷积块，当特征输入 Bottleneck 时，第一个卷积块核大小为 1 × 1，目的是降低输入的尺寸，减少数据量；中间通过卷积核为 3 × 3 的卷积，提取关键的特征图信息；最后通过第三个 1 × 1 卷积扩大特征的通道数，既保持了精度同时也减少了计算量。为了保证网络的整体稳定性，每个 Bottleneck 添加残差连接。

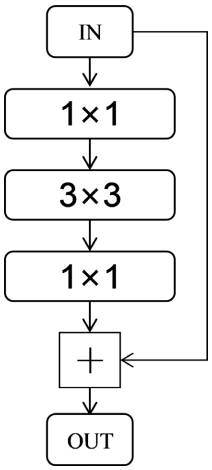

Figure 3. Bottleneck structure
图 3. Bottleneck 结构

本文采用的是 ResNet152 结构，该结构主要包括 5 层卷积层。其中第一个卷积层的卷积核为 7 × 7，后面的每一层都由 Bottleneck 搭建而成，具体的输入输出维度由表 1 所示。在本文中，我们使用 ImageNet 数据集进行预训练处理。

Table 1. ResNet152 dimension setting
表 1. ResNet152 维度设置

卷积层	输入维度	输出维度
Conv_1	(224, 224, 3)	(112, 112, 64)
Conv_2	(112, 112, 64)	(56, 56, 256)
Conv_3	(56, 56, 256)	(28, 28, 512)
Conv_4	(28, 28, 512)	(14, 14, 1024)
Conv_5	(14, 14, 1024)	(7, 7, 2048)

3.3. 特征移位卷积模块

为了使样本与样本之间共享特征，学习到与自身样本相近的特征并建立相关性，增强样本间的相互表征关系，本文设计了特征移位卷积模块，结构如图 4 所示。移位卷积包括一个移位操作和一个 1 × 1 的卷积操作。其中，移位操作的主要方法是将不同区域的特征移到当前区域，替换掉对应位置的特征，以此来实现特征的共享。1 × 1 的卷积操作是为了聚合同一通道内的特征，并且保持输入输出通道的一致性。定义一个固定的特征移位规则：给定特征结构 $F \in \mathbb{R}^{N\times C}$，对于第 i 个通道的平移距离 $d = i \bmod N$。

经过移位操作后，每个特征都能得到其它所有特征的信息。然而由于不同特征之间的关系强度不同，引入了一个可学习的权重参数 $M \in \mathbb{R}^{N\times N}$，其中 N 表示样本特征数，则移位卷积后的特征输出为：

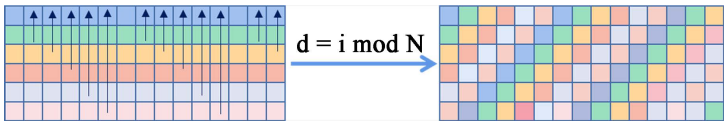

Figure 4. Feature shift convolution module
图 4. 特征移位卷积模块

$$F_{out} = \left\{ SCN\left(F_{in}\right) \odot M \right\} \odot F_{in} \tag{6}$$

经过移位之后的特征在对应某个样本能够同样学习到其它样本的特征，缩短同类动作样本向量在向量空间上的距离，增大异类样本向量的距离，进一步促进分类器对样本特征信息的学习。

移位卷积网络如图 5 所示。其中，特征图为上述两个网络处理后产生的高维特征，通过移位卷积后的新特征与权重参数 M 运算，并与原特征图进行残差连接，最后输入到全连接层，得到卷积后的结果。

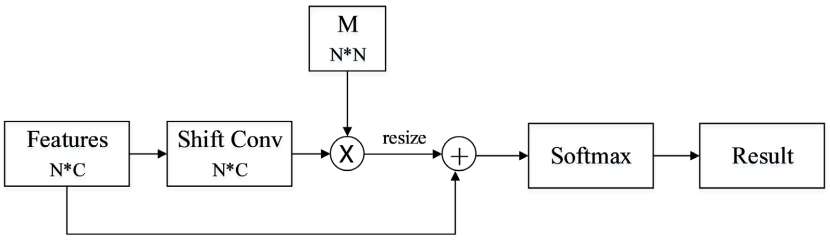

Figure 5. Shift convolution network module
图 5. 移位卷积网络模块

4. 实验结果与分析

4.1. 数据集

本文在大规模的第一人称手－动作(FPHA)数据集[28]上进行实验。该数据集是由 6 名操作者分别在 3 种不同的场景下演示拍摄，一共包含有 45 种动作类别，1175 个视频，总计 105,459 帧图像。在本文中，我们使用 600 个视频序列进行训练，其余的 575 个视频用于测试，并采用动作分类准确率进行评估。

4.2. 实验设置

在图卷积网络中，整个骨架提取网络包括 9 个自适应图卷积块，每一个自适应图卷积模块又包含空间域卷积和时间域卷积，两个卷积之后又分别连接一个 Dropout 层，丢失率设置为 0.5。为了保证网络训练的稳定，我们在每个图卷积模块添加残差连接。网络学习率设置为 0.1，批量大小设置为 32，一共设置 200 个 epoch。

在 ResNet152 网络训练过程中，我们使用交叉熵作为损失函数，选用动量值为 0.9 的随机下降梯度算法作为优化器，并且随着训练次数的增加相应调整学习率。

在经过上面两个网络提取得到的 RGB 特征和骨架特征，我们分别输入到特征移位卷积模块。其中，每组 RGB 特征输出维度较大，我们使用 1 × 1 的卷积层调整特征维度，然后进行特征的移位卷积操作，之后用 Softmax 全连接层进行分类的识别。在骨架特征中，由于每组训练数据的维度为 $N \times 256$，我们直接输入到特征移位卷积模块，之后用 Softmax 全连接层输出分类的分数。

4.3. 动作识别效果

为了验证本文所提出的网络模型的有效性，本文采用了不同数据模态，不同识别方法进行对比，实

验结果如表 2 所示。

Table 2. Action recognition performance comparison
表 2. 动作识别性能对比

模型	骨架序列	RGB 图像	准确率
Two stream [25]		√	75.30
H + O [26]		√	82.43
2s-AGCN [8]	√		84.12
Shift-GCN [27]	√		86.31
OURS	√	√	90.73

从表 2 可以看到，与目前主流的动作识别方法的比较，本文方法相比于其他方法可以取得更高的识别率，体现了本文框架的识别性能，相比之前提出的基准方法，本文方法可以取得 4.42%的提升。模型通过结合手部骨骼框架的方法，可以有效地处理视频图像中的背景噪声以及相机移动带来的干扰，提高神经网络模型的动作识别能力。除此之外，RGB 特征信息可以提供更加全面的场景信息进行分类；本文所使用的特征移位卷积模块可以使特征之间相互共享，对于最后的识别动作有很好的增益效果。值得注意的是，一个物体可能同时与多个动作相关联，因此我们将识别完全一致的动作组(动作 + 物体)，将其归为一次动作识别。该条件下，模型依旧展现出优异的识别能力。

4.4. 消融实验

为了验证框架种各网络结构的有效性，本文进行了消融实验。在该实验中，我们对网络中的各个模块进行单独测试与分析，并采用 Top-1 和 Top-5 准确度指标进行评估。

Table 3. Comparison of branch network performance
表 3. 分支网络性能对比

模型	Top-1	Top-5
AGCN + SCN	89.32	96.77
ResNet152 + SCN	81.74	97.83
All	90.73	98.61

我们首先测试本文网络框架中的网络分支对于整体性能的影响，得到的结果如表 3 所示。其中，AGCN (Adaptive graph convolution)代表单独使用 AGCN 网络分支处理骨架序列数据流，ResNet152 代表单独使用 ResNet152 分支网络处理 RGB 帧数据，SCN (Shift Convolution)表示移位卷积模块。从表中实验数据可以得到，融合两种数据流对于网络框架的整体准确率有明显的提升。具体分析，AGCN 在图结构方面的优异表现，偏向于对手骨架动作特征的提取和处理，而 ResNet152 偏向于对物体特征的识别，两者结合后有更强的增益效果。

参考文献

[1] Singh, S., Arora, C. and Jawahar, C.V. (2017) Trajectory Aligned Features for First Person Action Recognition. *Pattern Recognition*, **62**, 45-55. https://doi.org/10.1016/j.patcog.2016.07.031

[2] Kwon, H., Kim, Y., Lee, J.S. and Cho, M. (2018) First Person Action Recognition via Two-Stream ConvNet with Long-Term Fusion Pooling. *Pattern Recognition Letters*, **112**, 161-167. https://doi.org/10.1016/j.patrec.2018.07.011

[3] Lu, M., Li, Z.N., Wang, Y. and Pan, G. (2019) Deep Attention Network for Egocentric Action Recognition. *IEEE Transactions on Image Processing*, **28**, 3703-3713. https://doi.org/10.1109/TIP.2019.2901707

[4] Tang, Y., Wang, Z., Lu, J., Feng, J. and Zhou, J. (2018) Multi-Stream Deep Neural Networks for RGB-D Egocentric Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**, 3001-3015. https://doi.org/10.1109/TCSVT.2018.2875441

[5] Li, C., Xie, C., Zhang, B., Han, J., Zhen, X. and Chen, J. (2021) Memory Attention Networks for Skeleton-Based Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1-15. https://doi.org/10.1109/TNNLS.2021.3061115

[6] Du, Y., Fu, Y. and Wang, L. (2015) Skeleton Based Action Recognition with Convolutional Neural Network. 2015 3*rd IAPR Asian Conference on Pattern Recognition* (*ACPR*), Kuala Lumpur, 3-6 November 2015, 579-583. https://doi.org/10.1109/ACPR.2015.7486569

[7] Li, Y., He, Z., Ye, X., He, Z. and Han, K. (2019) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Dynamic Hand Gesture Recognition. *EURASIP Journal on Image and Video Processing*, **2019**, Article No. 78. https://doi.org/10.1186/s13640-019-0476-x

[8] Shi, L., Zhang, Y., Cheng, J. and Lu, H. (2019) Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 12018-12027. https://doi.org/10.1109/CVPR.2019.01230

[9] Su, K., Liu, X. and Shlizerman, E. (2020) PREDICT & CLUSTER: Unsupervised Skeleton Based Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 9628-9637. https://doi.org/10.1109/CVPR42600.2020.00965

[10] Singh, S., Arora, C. and Jawahar, C.V. (2016) First Person Action Recognition Using Deep Learned Descriptors. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2620-2628. https://doi.org/10.1109/CVPR.2016.287

[11] Urabe, S., Inoue, K. and Yoshioka, M. (2018) Cooking Activities Recognition in Egocentric Videos Using Combining 2Dcnn and 3Dcnn. *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*, Stockholm, 15 July 2018, 1-8. https://doi.org/10.1145/3230519.3230584

[12] Tang, Y., Wang, Z., Lu, J., Feng, J. and Zhou, J. (2018) Multi-Stream Deep Neural Networks for RGB-D Egocentric Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**, 3001-3015. https://doi.org/10.1109/TCSVT.2018.2875441

[13] Liu, Z., Zhang, H., Chen, Z., Wang, Z. and Ouyang, W. (2020) Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 140-149. https://doi.org/10.1109/CVPR42600.2020.00022

[14] Shi, L., Zhang, Y., Cheng, J. and Lu, H. (2020) Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *IEEE Transactions on Image Processing*, **29**, 9532-9545. https://doi.org/10.1109/TIP.2020.3028207

[15] Xia, H. and Gao, X. (2021) Multi-Scale Mixed Dense Graph Convolution Network for Skeleton-Based Action Recognition. *IEEE Access*, **9**, 36475-36484. https://doi.org/10.1109/ACCESS.2020.3049029

[16] Zhang, X., Xu, C. and Tao, D. (2020) Context Aware Graph Convolution for Skeleton-Based Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 14321-14330. https://doi.org/10.1109/CVPR42600.2020.01434

[17] Peng, W., Shi, J. and Zhao, G. (2021) Spatial Temporal Graph Deconvolutional Network for Skeleton-Based Human Action Recognition. *IEEE Signal Processing Letters*, **28**, 244-248. https://doi.org/10.1109/LSP.2021.3049691

[18] Cai, J., Jiang, N., Han, X., Jia, K. and Lu, J. (2021) JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 3-8 January 2021, 2734-2743. https://doi.org/10.1109/WACV48630.2021.00278

[19] Xie, J., Xin, W., Liu, R., Sheng, L., Liu, X., Gao, X., *et al*. (2021) Cross-Channel Graph Convolutional Networks for Skeleton-Based Action Recognition. *IEEE Access*, **9**, 9055-9065. https://doi.org/10.1109/ACCESS.2021.3049808

[20] Liu, J., Shahroudy, A., Xu, D., Kot, A.C. and Wang, G. (2018) Skeleton-Based Action Recognition Using Spatio-Temporal lstm Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 3007-3021. https://doi.org/10.1109/TPAMI.2017.2771306

[21] Nguyen, X.S., Brun, L., Lezoray, O. and Bougleux, S. (2019) A Neural Network Based on SPD Manifold Learning for Skeleton-Based Hand Gesture Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 12028-12037. https://doi.org/10.1109/CVPR.2019.01231

[22] Das, S., Sharma, S., Dai, R., Brémond, F. and Thonnat, M. (2020) VPN: Learning Video-Pose Embedding for Activities of Daily Living. *European Conference on Computer Vision* 2020, Glasgow, 23-28 August 2020, 72-90.

https://doi.org/10.1007/978-3-030-58545-7_5

[23] Yang, S., Liu, J., Lu, S., Er, M.H. and Kot, A.C. (2020) Collaborative Learning of Gesture Recognition and 3D Hand Pose Estimation with Multi-order Feature Analysis. *European Conference on Computer Vision* 2020, Glasgow, 23-28 August 2020, 769-786. https://doi.org/10.1007/978-3-030-58580-8_45

[24] He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[25] Feichtenhofer, C., Pinz, A. and Zisserman, A. (2016) Convolutional Two-Stream Network Fusion for Video Action Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 1933-1941. https://doi.org/10.1109/CVPR.2016.213

[26] Tekin, B., Bogo, F. and Pollefeys, M. (2019) H + O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 4506-4515. https://doi.org/10.1109/CVPR.2019.00464

[27] Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J. and Lu, H. (2020) Skeleton-Based Action Recognition with Shift Graph Convolutional Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 180-189. https://doi.org/10.1109/CVPR42600.2020.00026

[28] Garcia-Hernando, G., Yuan, S., Baek, S. and Kim, T.-K. (2018) First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations. *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 409-419. https://doi.org/10.1109/CVPR.2018.00050