一种视角无关的时空关联深度视频行为识别方法

吴培良[①②③] 杨 霄[①] 毛秉毅*[①③] 孔令富[①③] 侯增广[②]

[①](燕山大学信息科学与工程学院 秦皇岛 066004)

[②](中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190)

[③](河北省计算机虚拟技术与系统集成重点实验室 秦皇岛 066004)

摘 要：当前行为识别方法在不同视角下的识别准确率较低，该文提出一种视角无关的时空关联深度视频行为识别方法。首先，运用深度卷积神经网络的全连接层将不同视角下的人体姿态映射到与视角无关的高维空间，以构建空间域下深度行为视频的人体姿态模型(HPM)；其次，考虑视频序列帧之间的时空相关性，在每个神经元激活的时间序列中分段应用时间等级池化(RP)函数，实现对视频时间子序列的编码；然后，将傅里叶时间金字塔(FTP)算法作用于每一个池化后的时间序列，并加以连接产生最终的时空特征表示；最后，在不同数据集上，基于不同方法进行了行为识别分类测试。实验结果表明，该文方法(HPM+RP+FTP)提高了不同视角下深度视频识别准确率，在UWA3DII数据集中，比现有最好方法高出18%。此外，该文方法具有较好的泛化性能，在MSR Daily Activity3D数据集上得到82.5%的准确率。

关键词：视频行为识别；深度视频；视角无关；卷积神经网络；时空关联

中图分类号：TP242.6+2 **文献标识码**：A **文章编号**：1009-5896(2019)04-0904-07

DOI: 10.11999/JEIT180477

A Perspective-independent Method for Behavior Recognition in Depth Video via Temporal-spatial Correlating

WU Peiliang[①②③] YANG Xiao[①] MAO Bingyi[①③]

KONG Lingfu[①③] HOU Zengguang[②]

[①](*School of Information Science and Technology*, *Yanshan University*, *Qinhuangdao* 066004, *China*)

[②](*State Key Laboratory of Management and Control for Complex Systems*, *Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

[③](*The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province*, *Qinhuangdao* 066004, *China*)

Abstract: Considering the low recognition accuracy of behavior recognition from different perspectives at present, this paper presents a perspective-independent method for depth videos. Firstly, the fully connected layer of depth Convolution Neural Network (CNN) is creatively used to map human posture in different perspectives to high-dimensional space that is independent with perspective to achieve the Human Posture Modeling (HPM) of deep-performance video in spatial domain. Secondly, considering temporal-spatial correlation between video sequence frames, the Rank Pooling (RP) function is applied to the series of each neuron activated time to encode the video time sub-sequence, and then the Fourier Time Pyramid (FTP) is used to each pooled time series to produce the final spatio-temporal feature representation. Finally, different methods of behavior recognition classification are tested on several datasets. Experimental results show that the proposed method improves the accuracy of depth video recognition in different perspectives. In the UWA3DII datasets, the proposed method is 18% higher than the most recent method. The proposed method

收稿日期：2018-05-21；改回日期：2018-12-04；网络出版：2018-12-14

*通信作者： 毛秉毅 ysdxmby@163.com

基金项目：国家自然科学基金(61305113)，河北省自然科学基金(F2016203358)，中国博士后基金(2018M631620)，燕山大学博士基金(BL18007)

Foundation Items: The National Natural Science Foundation of China (61305113), The Natural Science Foundation of Hebei Province (F2016203358), China Postdoctoral Science Foundation (2018M631620), The Doctoral Fund of Yanshan University (BL18007)

(HPM+RP+FTP) has a good generalization performance, achieving a 82.5% accuracy on dataset of MSR Daily Activity3D.

Key words: Video behavior recognition; Depth video; Perspective-independent; Convolution Neural Network (CNN); Temporal-spatial correlation

1　引言

视频中人体行为识别在智能监控、智能家居等环境下具有重要应用前景。然而，受摄像机视角影响，基于多视角视频下的人体行为识别研究具有较高的挑战性。在行为识别领域，根据视频类型不同，可将主流方法归纳为2类：基于RGB视频的行为识别和基于深度视频的行为识别。在基于RGB视频的行为识别方面，文献[1]提出了一种基于细粒度的行为识别方法，该方法采用判别式挖掘算法，在单一视角下的RGB视频中识别准确率良好，然而当视角发生变化后，该方法的识别准确率就会急剧下降。针对该问题，文献[2]考虑不同视角之间时空模型和几何关系，通过单独学习不同视角之间的每个身体部位的线性变换来训练时空的与或图结构，然后采用穷举法获得识别结果，但该方法受局部特征选择的影响。在基于深度视频行为识别方面，文献[3]提出了HON4D方法。文献[4]提出通过连接每个像素点间局部邻域的4D法线作为描述符来扩展HON4D方法，然而，上述描述符必须在可靠兴趣点上提取。为了解决这一问题，文献[5]提出了一种滤除深度传感器噪声并提取更可靠时空兴趣点的方法，但该方法对视角变化下的行为视频识别精度较低。

此外，基于人体骨架结构的行为识别也比较常见。文献[6]从视频的每一帧中提取人体关节点固定区域的直方图特征，提出了一种识别行为类别中最具歧义性关节点的数据挖掘算法。文献[7]基于人体骨架表示来建立相对几何模型的李群曲线，但在非正视视角下人体关节点提取精度较低，造成行为识别错误率增加。

近年来，深度学习渐为流行。文献[8]较早使用卷积神经网络，但是性能却较传统的手工设计特征提取方法低[9]。文献[10]利用光流法设计了包含空间和时间网络的双流CNN。文献[11]提出了3D CNN方法。文献[12]利用递归神经网络(RNN)对动作进行动态编码。上述几种深度学习方法均基于RGB视频进行行为识别，训练这些模型需要大量的带标签样本，且当识别其他数据集时需对原有模型进行再训练或微调。

文献[8]发现：深度卷积神经网络的单帧模型同样适用于多帧模型。受此启发，本文利用多视角人体姿态图像样本训练深度卷积神经网络，然后将该深度卷积神经网络应用于多视角深度视频中，创造性地运用深度卷积神经网络的全连接层将不同视角下的人体姿态映射到与视角无关的高维空间，形成人体姿态模型。并考虑到视频的时空关联特性，利用时间等级池化和傅里叶时间金字塔的方法对时间结构进行处理，形成最终时空联合表示。实验表明，本文方法对不同视角下深度视频行为的识别率大幅提高，同时，本文方法具有较好的泛化性能，在识别其他数据集时并不需要重新对模型进行训练或微调。

2　模型整体框架及描述

整体模型框架分为5个部分，如图1所示，第1部分为输入的数据，即需要识别的连续深度行为视频帧；第2部分为经过训练调整的CNN模型，输入的深度行为视频帧经过CNN后输出全连接层的4096维特征向量，即所需识别视频的与视角无关的空间特征向量；第3部分为时间等级池化函数，对视频随时间变化的外观演化进行编码表示；第4部分为傅里叶时间金字塔，用以处理经过时间等级池化函数后的特征向量，从而对视频时间结构进行建模并对特征向量做暂时对齐处理；第5部分为支持向量机分类器，对特征向量进行分类，得到最终的识别准确率。

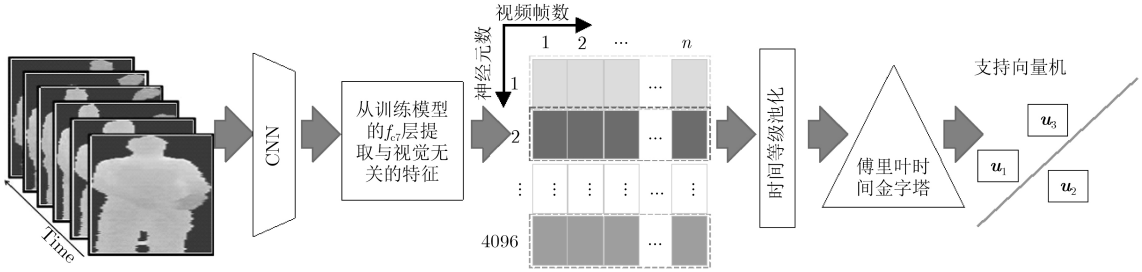

图 1 整体模型框架

3 人体姿态估计模型

本文所构建的与视角无关的人体姿态模型是一个深度卷积神经网络，该模型的结构类似于文献[13]，但考虑到训练数据应用的多视角人体姿态图像的动作类别只有339类，故将文献[13]深度卷积神经网络模型中最后的全连接层置换为含有339个神经元的全连接层，并在训练阶段，于CNN网络之后加入softmax损失层实现训练样本的分类输出，此外，为防止网络过拟合，在CNN网络中加入2层辍学率为0.5的辍学层。本文采用的CNN模型整体结构如图2所示，具体训练过程将在后面进行详细表述。本模型应用的最大值池化的步长均为2，激活函数均为relu函数。将本文模型的3个全连接层从左至右分别记为f_{c6},f_{c7},f_{c8}，则f_{c6}层和f_{c7}层实现了将不同视角的人体姿态映射到与视角无关的高维空间，含有339个神经元的f_{c8}层后面加上softmax损失层实现了训练过程中339个动作类别的输出。

本文中，训练好的CNN模型的输入是人体姿态的深度图像，输出是对应的姿态类别。为了使该模型从深度视频中提取视角不变的特征，首先进行下面2个步骤。

(1)数据预处理：训练CNN模型用的数据集中只包含人类姿势的深度图像。因此，需要把人体姿态从整体背景中分割出来。Kinect摄像机可以完成分割与裁剪，形成与CNN模型相兼容的形式。具体而言，将从背景中分离出的人体姿态深度图像的像素值标准化到0～255的范围，大小转变为227×227的尺寸，再将待处理图像各像素值减去平均深度值。在人体未被分割时，即含有复杂背景的情况下，该模型也有较好的识别效果。

(2)特征提取：对于每个深度视频，经过预处理的视频帧图像通过CNN网络后，利用f_{c7}层输出的向量作为视角不变的行为特征描述符，实验部分证明了f_{c7}层输出向量比f_{c6}层的识别精度更高。

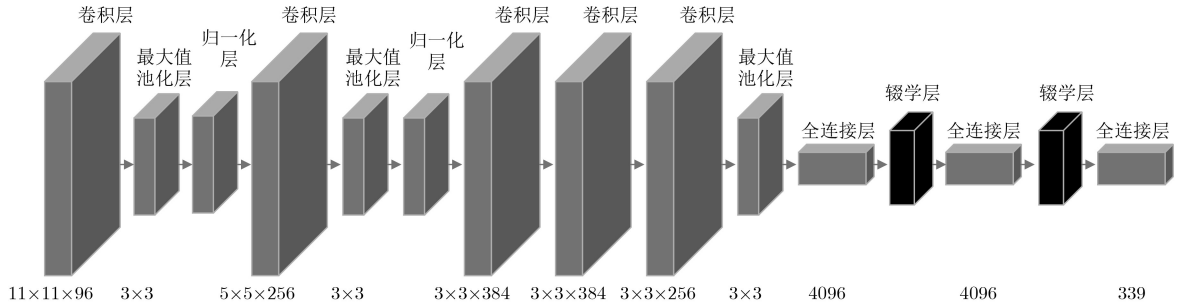

图 2 本文采用的CNN模型结构

4 时间等级池化

由于视频是具有时间信息的帧序列，本文在每个神经元激活的时间序列中分段应用时间等级池化函数[14]，实现对视频时间子序列的编码。该方法仅处理帧与帧之间的相对位置关系，对视频速度不敏感。

4.1 时间表示的函数参数

假设视频的第i帧由向量$\boldsymbol{x}_i$表示，那么由n帧组成的视频就可以由向量序列$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\cdots\boldsymbol{x}_n]$来表示。首先对经过CNN处理的视频利用中值滤波进行平滑处理，从而得到一个新的向量序列$\boldsymbol{V}=[\boldsymbol{v}_1\ \boldsymbol{v}_2\cdots\boldsymbol{v}_n]$。时间等级池化是对视频外观随时间的演变过程进行编码，也就是对序列$\boldsymbol{V}$进行动态编码得到序列$\boldsymbol{D}$，从抽象角度来看，动态编码序列$\boldsymbol{D}$表示了输入向量从时间t到$t+1$的所有变化。假设向量$\boldsymbol{V}$足够平滑，可以用一个动态函数$\psi_u=\psi(\boldsymbol{V};u)$中的参数$u$来对向量$\boldsymbol{V}$进行动态编码，即使得$\psi$逼近$\boldsymbol{D}$，如式(1)所示：

$$\arg\min_{u}\|\boldsymbol{D}-\psi\| \tag{1}$$

由于相同动作类别的不同视频在外观的动态表示上是不尽相同的，而动态编码序列$\boldsymbol{D}$中包含组合函数ψ，对于每一个视频$\boldsymbol{V}_i(\cdot)$，学习以$\boldsymbol{u}_i$为因变量的不同动态函数$\psi_i(\cdot\,;\boldsymbol{u}_i)$，进而得到稳定且鲁棒的组合函数$\psi$。其中函数变量$\boldsymbol{u}_i$作为视频新的表示，这样就得到了用于时间表示的函数参数$\boldsymbol{u}_i$。

4.2 等级池化方法

视频是有序的帧序列，其中帧序列的顺序决定视频的外观演化，如果$\boldsymbol{v}_{t+1}$排列在$\boldsymbol{v}_t$之后，就可以得到$\boldsymbol{v}_{t+1}\succ\boldsymbol{v}_t$的排序表示。同理，可以得到最终的约束条件为$\boldsymbol{v}_n\succ\cdots\succ\boldsymbol{v}_t\succ\cdots\succ\boldsymbol{v}_1$。本文利用视频帧的传递性制定目标，即逐对式的等级排名机制，$\boldsymbol{v}_a\succ\boldsymbol{v}_b,\ \boldsymbol{v}_b\succ\boldsymbol{v}_c\Rightarrow\boldsymbol{v}_a\succ\boldsymbol{v}_c$。

使用该等级排名机制对视频进行动态建模，本质上是解决约束最小化的问题。解决方法是利用线性等级排名机制学习线性函数$\psi(\boldsymbol{v};\boldsymbol{u})=\boldsymbol{u}^{\mathrm{T}}\cdot\boldsymbol{v}$，其中$\boldsymbol{u}\in R^{\mathrm{D}}$。为了避免过拟合，$\boldsymbol{v}_t$的等级排名是由$\psi(\boldsymbol{v}_t;\boldsymbol{u})=\boldsymbol{u}^{\mathrm{T}}\cdot\boldsymbol{v}_t$和满足逐对式约束$(\boldsymbol{v}_{t+1}\succ\boldsymbol{v}_t)$的边界参数决定。因此，需要学习满足所有约束条

件($\forall t_i, t_j$, $\boldsymbol{v}_{t_i} \succ \boldsymbol{v}_{t_j} \Leftrightarrow \boldsymbol{u}^{\mathrm{T}} \cdot \boldsymbol{v}_{t_i} > \boldsymbol{u}^{\mathrm{T}} \cdot \boldsymbol{v}_{t_j}$)的参数向量$\boldsymbol{u}$。利用结构风险最小化和结构边缘最大化的机制，约束等级池化的目标函数为

$$\left.\begin{aligned} &\arg\min_{u} \frac{1}{2}\|\boldsymbol{u}\|^2 + C \sum_{\forall i,j \boldsymbol{v}_{t_i} \succ \boldsymbol{v}_{t_j}} \in_{ij} \\ &\text{s.t. } \boldsymbol{u}^{\mathrm{T}} \cdot \left(\boldsymbol{v}_{t_i} - \boldsymbol{v}_{t_j}\right) \geq 1 - \in_{ij}, \in_{ij \geq 0} \end{aligned}\right\} \tag{2}$$

然后，利用支持向量回归(SVR)来估计这个线性函数并学习参数$\boldsymbol{u}$，对于每一个视频子序列，利用学习到的参数$\boldsymbol{u}$表示经过池化后的特征。

通过时间等级池化的方法建模有以下几个优点：首先，在视频中一些动作的速度变化过于迅速，该方法对视频速度并不敏感。其次，时间等级池化的方法对CNN特征，词袋模型的向量和费舍尔编码矢量均有很好的处理能力。因此本文利用时间等级池化的方法对经过CNN模型激活的视频帧进行子序列动态编码。

5　全局时间编码和分类

一些深度视频是由低成本的深度摄像机拍摄得到，这些数据往往会含有较高噪声，此外，不同的视频帧长不一致，经过时间等级池化函数处理后的视频子序列的特征向量组合在数据维度上参差不齐，强行改变数据维度会丢失特征信息，造成识别精度出现一定程度的降低(实验部分已加以验证)。鉴于此，需要一个对含噪声深度视频鲁棒且可使视频特征向量暂时对齐的方法。

傅里叶时间金字塔(FTP)[15]被证明能简单而鲁棒地处理含有噪声的时间序列的方法，并且能够使特征向量暂时对齐。首先设定全局傅里叶级数，然后递归地分割时间等级池化后的视频子序列的特征向量并对所有序列的特征向量进行快速傅里叶变换，最后串联所有的傅里叶系数来表示完整的时间结构描述符。

在$F = 1,2,\cdots,L$层的金字塔中，令$\theta^i_{k,l}$表示第i个视频的子序列的动态特征，$\theta^i_{k,l}$中的$k = 1,2,\cdots,D$表示特征数目的索引，$l = 1,2,\cdots,N$表示帧数。在最低的金字塔层，执行对$\boldsymbol{\theta}^i_k = \left[\theta^i_{k,1},\theta^i_{k,2},\cdots,\theta^i_{k,N}\right]$的FFT，得到第1组低频系数$p$，保持第1组低频系数不变，将$\theta^i_k$分为2部分，将该方法应用在每一部分，如此迭代重复，直到迭代次数达到预设的傅里叶时间金字塔的层数后停止迭代。所有低频系数的连接被用作每一个条目的特征，最后，将所有条目的FTP特征组合起来作为第i个视频的时空表示，这样就得了一个视频最终的特征向量。

在分类部分，采用支持向量机作为动作分类器来预测动作的分类，支持向量机在动作识别领域是应用最为广泛并且效果最好的分类器之一。在本文提出的方法中，只需要在最后的动作分类器中用到监督学习的机制，故可有效避免重复训练。

6　实验结果分析

为评估本文方法的性能，在一个交叉视角的数据集UWA 3DII和一个背景复杂单一视角的数据集MSR Daily Activity3D上进行测试。在实验中，设定时间等级池化方法的步长为2，视频帧分段间隔为15帧，傅里叶金字塔的层数为4层，第1级低频傅里叶级数为4。出于简化，将上文提到人体姿态模型记为HPM，时间等级池化方法记为RP，傅里叶时间金字塔记为FTP，将本方法与现有方法进行比较。

6.1　数据集

6.1.1　训练数据集

本文训练用的数据集是UWA大学基于CMU数据集制作的[16]。CMU数据集由高精度的摄像机拍摄获得，其中至少包含2×10^5个人体姿态，但这些姿态有一些不具有代表性的重复部分，所以首先利用k-means方法从CMU数据集中进行聚类，然后随机地选取339个最具代表性的人体姿态，如挥手、下蹲、静坐等。再利用开源的MakeHuman软件、Blender工具包和180个呈现半球位置的虚拟摄像机将选取好的人体姿态数据制作成多视角的深度图像数据集。最终得到339类人体动作姿态图像，每个动作拥有180个视角。该数据只有人体姿态部分，并不包含复杂的背景，图像尺寸均为227×227，像素值均在0～255之间。选用该数据集是因为所选取的339个动作类别具有代表性，而且180个视角可以包含绝大部分视角位置，这一特点是训练出与视角无关的深度卷积神经网络的基础。

6.1.2　UWA3DII数据集

该数据集包含30个动作类别：(1)单手挥手、(2)单手打拳、(3)双手挥手、(4)双手打拳、(5)坐下、(6)起立、(7)摇晃、(8)跌倒、(9)抚摸前胸、(10)抚摸头部、(11)抚摸后背、(12)走路、(13)不规则走路、(14)躺下、(15)转身、(16)喝水、(17)打电话、(18)俯身、(19)曲腿挥手、(20)奔跑、(21)捡起东西、(22)放下东西、(23)踢、(24)跳跃、(25)跳舞、(26)呕吐、(27)打喷嚏、(28)静坐、(29)蹲下、(30)咳嗽。每一个动作表演者做4次，每一次均由顶部，正面，左侧，右侧4个角度进行拍摄。该视频的挑战性在于视频是不同角度拍摄的，这造成很大程度的遮挡，比如在顶部的视角中，身体的下半部分就因为遮挡没有得到适当的捕捉。而且数据集

中动作类别比较多，一些动作的相似程度较大，也对动作的识别提出挑战。

6.1.3 MSR Daily Activity3D数据集

MSR Daily Activity3D数据集是由Kinect深度摄像机拍摄的16种日常行为动作组成的，其中包括：(1)喝水、(2)吃东西、(3)看书、(4)打电话、(5)写字、(6)使用笔记本电脑、(7)使用吸尘器、(8)欢呼雀跃、(9)静坐、(10)扔纸团、(11)做游戏、(12)躺在沙发上、(13)走路、(14)弹吉他、(15)起立、(16)坐下。这些动作分别由表演者以两种方式完成，一种是坐在沙发上表演，另一种为站立着表演，该数据集包含320个样本。

6.2 实验条件及参数设置

因为训练数据集由180个虚拟相机以环绕半球的方位拍摄，本文利用随机选取的162个角度的深度图像进行训练，剩余的18个角度的深度图像进行测试。正确的初始化是一个CNN模型成功的关键，并且要避免过拟合的问题发生。本文选取2012年ImageNet挑战赛上训练了1.2×10^6张RGB图像并且经过NYUD2深度图像数据集微调的CNN模型[13]，将该模型调整为第2节所述的形式，利用反向传播算法训练，超参数设置分别为：卷积层和全连接层的学习率设置为0.01，冲量单元设置为0.9，权重衰减率为0.0005。设置2.1×10^4次迭代来训练模型，在训练过程中，利用图像翻转的方式进行数据增广，输入图像的水平翻转概率为0.5。

6.3 UWA3DII数据集实验验证

选取2个视角的数据集作为训练集，其余2个视角的数据集作为测试集。表1展示了本文方法与其他方法的结果对比。文献[17]的准确率相对较高，但是文献[17]算法必须要人工设计提取密集轨迹特征。利用本文的HPM+FTP的方法，评价识别率达到了76.9%，比文献[17]的方法提高了13.4%，但仍存在一些动作的识别不够准确，当加入RP的方法对视频的外观演化进行处理之后，识别率再次提高了1.4%，达到78.3%。此外需要指出的是，由于视频帧数的不一致性导致RP处理后的特征维度不一致，直接改变维度分类的平均准确率为71.3%，比HPM+RP+FTP的方法低7%，对比可见FTP方法的加入在对齐特征的同时有效地保证了识别精度。值得注意的是本文的方法在用顶部视角进行测试的时候平均准确率达到77.9%，因为顶部视角下某些动作下肢部分是被遮挡的，故可表明本文方法对有遮挡的视频识别具有鲁棒性。

表 1 UWA3D Multiview ActivityII数据集的动作识别准确性(%)

训练视角	V1&V2		V1&V3		V1&V4		V2&V3		V2&V4		V3&V4		平均准确率
测试视角	V3	V4	V2	V4	V2	V3	V1	V4	V1	V3	V1	V2	
文献[6]	45.0	40.4	35.1	36.9	34.7	36.0	49.5	29.3	57.1	35.4	49.0	29.3	39.8
文献[7]	49.4	42.8	34.6	39.7	38.1	44.8	53.3	33.5	53.6	41.2	56.7	32.6	43.4
文献[18]	52.7	51.8	59.0	57.5	42.8	44.2	58.1	38.4	63.2	43.8	66.3	48.0	52.2
文献[17]	60.1	61.3	57.1	65.1	61.6	66.8	70.6	59.5	73.2	59.3	72.5	54.5	63.5
HPM(f_{c7})+RP	80.2	74.9	69.9	76.4	49.2	63.8	71.4	59.9	80.7	76.9	84.4	68.4	71.3
HPM(f_{c7})+FTP	80.6	80.5	75.2	**82.0**	65.4	72.0	77.3	67.0	**83.6**	81.0	83.6	74.1	76.9
HPM(f_{c6})+RP+FTP	83.9	81.3	74.8	**82.0**	**66.2**	72.8	**78.8**	70.0	83.3	79.1	**85.9**	75.9	77.8
HPM(f_{c7})+RP+FTP	**85.8**	**81.6**	**76.3**	80.5	61.7	**76.5**	78.1	**71.5**	82.9	**81.7**	**85.9**	**76.3**	**78.3**

注：V_1, V_2, V_3, V_4分别表示正面视角、左侧视角、右侧视角、顶部视角

图3展示了本文提出的HPM+FTP和HPM+RP+FTP方法对于几种特定动作的识别率，可以看出HPM+RP+FTP方法对于双手挥手、摇晃、喝水、跳跃有明显的提高，这是由于这几种动作的外观随时间演变较为明显。但是也有一些动作的识别准确率有轻微下降如坐下等，这可能是由于在视频分段进行等级池化时有些连续动作部分被分割开造成的，还有待进一步的研究。

值得强调的是，在UWA3DII数据集中的一些动作，如双手挥手、抚摸头部、抚摸后背、打喷嚏和咳嗽，在训练CNN模型时所用的CMU mocap数据集中并不存在，所以该模型对于不同姿态也具有较强泛化能力。

6.4 MSR Daily Activity3D数据集实验验证

该数据集涵盖了人类在客厅里的日常活动。当表演者站在靠近沙发或是坐在沙发上时，会对Kinect深度摄像机采集数据时产生过多的噪声，对于动作的识别造成很大的影响。此外，该数据的大多数动作都涉及到人物的交互，本文使用的CNN模型并没有利用类似的数据进行训练或微调，为了验证本文模型的泛化能力，采用该数据集作为第2个实验。

在表2中展示了几种方法对该数据集的识别准

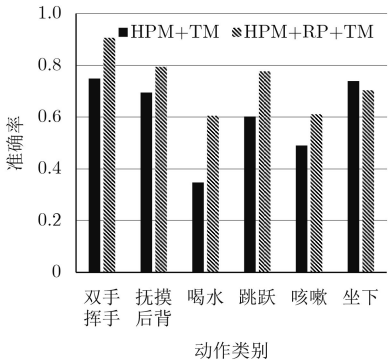

图 3 比较2种方法对于特点动作的识别准确率

表 2　几种方法对MSR Daily Activity3D的准确率(%)

方法	准确率
文献[19]	79.1
文献[20]	54.0
文献[21]	68.0
文献[22]	73.8
HPM(fc7)+RP	60.0
HPM(fc7)+FTP	79.9
HPM(fc6)+RP+FTP	81.3
HPM(fc7)+RP+FTP	**82.5**

确率，在实验的分类部分，利用160条深度视频做训练(subject=1, 3, 5, 7, 9)，其余160条做测试，可以看到本文提出的HPM+FTP的方法在该数据上得到了79.9%的准确率，比其余几种方法的准确率相比均有提升，但是只比文献[19]的方法高出0.8%，接下来对视频的外观随时间的演化进行RP的进一步处理，得到了82.5%的准确率，充分验证了本文所提出的方法对于各类行为识别的泛化能力。若未利用FTP方法进行时间结构的处理，在此泛化能力验证的数据上准确率降低22.5%，进一步验证了FTP方法的重要性。图4展示了这16种动作的识别准确率的混淆矩阵图，在混淆矩阵中可以看到一些与物体交互，特别是在沙发附近做动作可能受到干扰的动作，识别准确率还有待进一步提高。

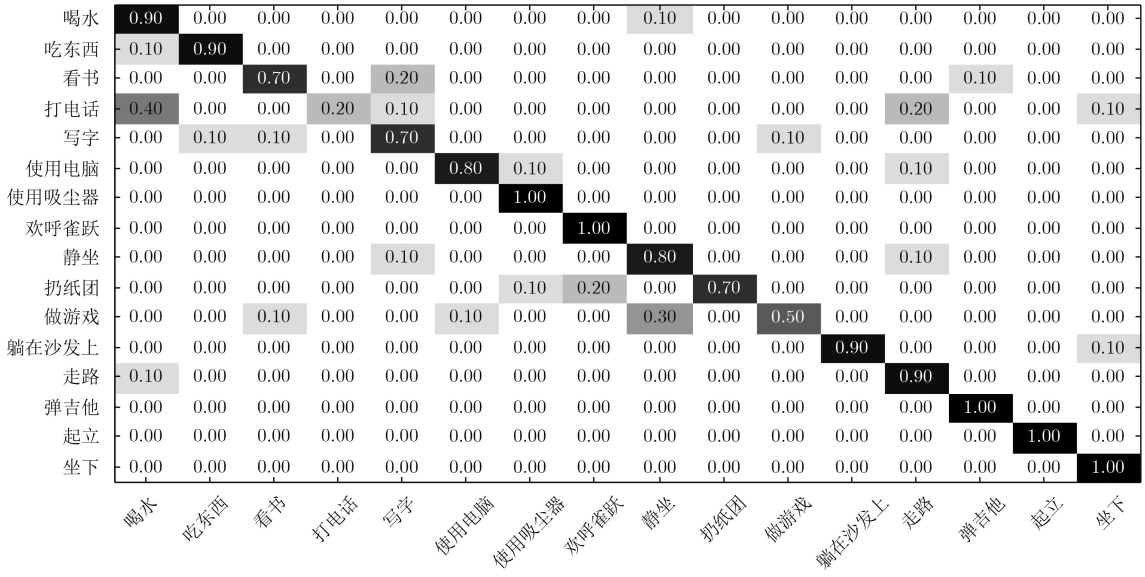

图 4 MSR Daily Activity3D数据集16种动作的混淆矩阵

7　结论

本文设计了一种深度CNN人体姿态模型HPM，用于从多视角获得的人体姿态深度图像中提取高维不变空间特征，此外，提出了一种应用于视频时间结构建模的时间等级池化方法，并将其与傅里叶时间金字塔相结合，形成对深度动作视频数据的时空表示。在交叉视角的深度视频数据集以及背景复杂的单一视角数据集上，本文所提出的HPM+RP+FTP方法均取得了理想的识别效果。

参 考 文 献

[1] ZHOU Yang, NI Bingbing, HONG Richang, *et al*. Interaction part mining: A mid-level approach for fine-grained action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3323–3331. doi: 10.1109/CVPR.2015.7298953.

[2] WANG Jiang, NIE Xiaohan, XIA Yin, *et al*. Cross-view action modeling, learning, and recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 2649–2656. doi: 10.1109/CVPR.2014.339.

[3] LIU Peng and YIN Lijun. Spontaneous thermal facial expression analysis based on trajectory-pooled fisher vector descriptor[C]. IEEE International Conference on Multimedia and Expo, Hong Kong, China, 2017: 835–840. doi: 10.1109/

ICME.2017.8019315.

[4] YANG Xiaodong and TIAN Yingli. Super normal vector for activity recognition using depth sequences[C]. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 804–811. doi: 10.1109/CVPR.2014.108.

[5] ZHANG Baochang, YANG Yun, CHEN Chen, *et al*. Action recognition using 3D histograms of texture and a multi-class boosting classifier[J]. *IEEE Transactions on Image Processing*, 2017, 26(10): 4648–4660. doi: 10.1109/TIP.2017.2718189.

[6] YIN Xiaochuan and CHEN Qijun. Deep metric learning autoencoder for nonlinear temporal alignment of human motion[C]. IEEE International Conference on Robotics and Automation, Stockholm, Sweden, 2016: 2160–2166. doi: 10.1109/ICRA.2016.7487366.

[7] SHAHROUDY A, LIU Jun, NG T, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.

[8] KARPATHY A, TODERICI G, SHETTY S, *et al*. Large-scale video classification with convolutional neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1725–1732. doi: 10.1109/CVPR.2014.223.

[9] HAIDER F, CAMPBELL N, and LUZ S. Active speaker detection in human machine multiparty dialogue using visual prosody information[C]. IEEE Global Conference on Signal and Information Processing, Washington, D.C., USA, 2016: 1207–1211. doi: 10.1109/GlobalSIP.2016.7906033.

[10] SIMONYAN K and ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. *Advances in Neural Information Processing Systems*, 2014, 1(4): 568–576. doi: 10.1002/14651858.CD001941.pub3.

[11] TRAN D, BOURDEV L, FERGUS R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C]. IEEE International Conference on Computer Vision, Honolulu, USA, 2015: 4489–4497. doi: 10.1109/ICCV.2015.510.

[12] DONAHUE J, HENDRICKS L A, ROHRBACH M, *et al*. Long-term recurrent convolutional networks for visual recognition and description[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 677–691. doi: 10.1109/TPAMI.2016.2599174.

[13] GUPTA S, GIRSHICK R, ARBELEZ P, *et al*. Learning rich features from RGB-D images for object detection and segmentation[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014: 345–360. doi: 10.1007/978-3-319-10584-0_23.

[14] FERNANDO B, GAVVES E, ORAMAS J, *et al*. Rank pooling for action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 773–787. doi: 10.1109/TPAMI.2016.2558148.

[15] WANG Jiang, LIU Zicheng, WU Ying, *et al*. Learning actionlet ensemble for 3D human action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(5): 914–927. doi: 10.1109/TPAMI.2013.198.

[16] RAHMANI H and MIAN A. 3D action recognition from novel viewpoints[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1506–1515. doi: 10.1109/CVPR.2016.167.

[17] RAHMANI H and MIAN A. Learning a non-linear knowledge transfer model for cross-view action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 2458–2466. doi: 10.1109/CVPR.2015.7298860.

[18] RAHMANI H, MAHMOOD A, HUYNH D Q, *et al*. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014: 742–757. doi: 10.1007/978-3-319-10605-2_48.

[19] JALAL A, KAMAL S, and KIM D. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments[J]. *Sensors*, 2014, 14(7): 11735–11759. doi: 10.3390/s140711735.

[20] MULLER M and RODER T. Motion templates for automatic classification and retrieval of motion capture data[C]. ACM Siggraph/eurographics Symposium on Computer Animation, Vienna, Austria, 2006: 137–146. doi: 10.1145/1218064.1218083.

[21] WANG Jiang, LIU Zicheng, WU Ying, *et al*. Mining actionlet ensemble for action recognition with depth cameras[C]. IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 1290–1297. doi: 10.1007/978-3-319-04561-0_2.

[22] CAVAZZA J, ZUNINO A, BIAGIO M S, *et al*. Kernelized covariance for action recognition[C]. International Conference on Pattern Recognition, Cancun, Mexico, 2016: 408–413. doi: 10.1109/ICPR.2016.7899668.

吴培良：男，1981年生，副教授，研究方向为家庭服务机器人行为识别与学习、功用性认知.

杨 霄：男，1993年生，硕士生，研究方向为家庭服务机器人行为识别.

毛秉毅：男，1964年生，副研究员，研究方向为家庭服务机器人.

孔令富：男，1957年生，教授，研究方向为智能机器人系统、智能信息处理.

侯增广：男，1969年生，研究员，研究方向为机器人与智能系统、康复机器人与微创介入手术机器人.