E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2024)04-1056-14

论文引用格式:You K J, Hou Z J, Liang J Z, Zhong Z K and Shi H Y. 2024. Point cloud human behavior recognition based on coordinate transformation and spatiotemporal information injection. Journal of Image and Graphics, 29(04):1056-1069(尤凯军, 侯振杰, 梁久祯, 钟卓锟, 施海勇. 2024. 结合坐标转换和时空信息注入的点云人体行为识别. 中国图象图形学报, 29(04):1056-1069)[DOI:10.11834/jig.230215]

结合坐标转换和时空信息注入的点云人体行为识别

尤凯军,侯振杰*,梁久祯,钟卓锟,施海勇

常州大学计算机与人工智能学院,常州　213000

摘　要:**目的**　行为识别中广泛使用的深度图序列存在着行为数据时空结构信息体现不足、易受深色物体等因素影响的缺点,点云数据可以提供丰富的空间信息与几何特征,弥补了深度图像的不足,但多数点云数据集规模较小且没有时序信息。为了提高时空结构信息的利用率,本文提出了结合坐标转换和时空信息注入的点云人体行为识别网络。**方法**　通过将深度图序列转换为三维点云序列,弥补了点云数据集规模较小的缺点,并加入帧的时序概念。本文网络由两个模块组成,即特征提取模块和时空信息注入模块。特征提取模块提取点云深层次的外观轮廓特征。时空信息注入模块为轮廓特征注入时序信息,并通过一组随机张量投影继续注入空间结构信息。最后,将不同层次的多个特征进行聚合,输入到分类器中进行分类。**结果**　在3个公共数据集上对本文方法进行了验证,提出的网络结构展现出了良好的性能。其中,在NTU RGB+d60数据集上的精度分别比PSTNet(point spatio-temporal network)和SequentialPointNet提升了1.3%和0.2%,在NTU RGB+d120数据集上的精度比PSTNet提升了1.9%。为了确保网络模型的鲁棒性,在MSR Action3D小数据集上进行实验对比,识别精度比SequentialPointNet提升了1.07%。**结论**　提出的网络在获取静态的点云外观轮廓特征的同时,融入了动态的时空信息,弥补了特征提取时下采样导致的时空损失。

关键词:人体行为识别;坐标转换;点云序列;特征提取;时空信息

Point cloud human behavior recognition based on coordinate transformation and spatiotemporal information injection

You Kaijun, Hou Zhenjie*, Liang Jiuzhen, Zhong Zhuokun, Shi Haiyong

College of Computer and Artificial Intelligence, Changzhou University, Changzhou　213000, China

Abstract: **Objective** Human motion recognition and deep learning have become a research hotspot in the field of computer vision because of their extensive applications in video surveillance, virtual reality, and human computer intelligent interaction. Deep learning theory has made excellent achievements in the feature extraction of static images and has been gradually extended to the research of behavior recognition in other directions. Traditional research on human behavior recognition focuses on depth image sequence under 2D information. Depth image cannot only capture 3D information successfully, but can also provide depth information. Depth information represents the distance between the target and the depth camera within the visual range, disregarding the influence of external factors, such as lighting and background. Although depth image can capture 3D information, most depth image algorithms use the multi-view method to extract behavior features.

收稿日期:2023-04-15;**修回日期**:2023-07-03;**预印本日期**:2023-07-10

*通信作者:侯振杰　houzj@cczu.edu.cn

基金项目:国家自然科学基金项目(61063021);江苏省研究生科研创新计划(KYCX21_2834, KYCX21_2835)

Supported by: National Natural Science Foundation of China (61063021)

The extraction effect of spatiotemporal features is affected by the angle and number of multiple views, considerably affecting the utilization rate of 3D structural information, and the spatiotemporal structure information of 3D data is largely lost. With the rapid development of 3D acquisition technology, 3D sensors are becoming increasingly accessible and affordable, including various types of 3D scanners and LiDAR. The 3D data collected by these sensors can provide rich geometry, shape, and scale information. 3D data have many applications in different fields, including autonomous driving, robotics, remote sensing, and healthcare. Point cloud representation is a commonly used 3D representation; it retains the original geometric information in 3D space without any discretization. Therefore, it is the preferred representation for understanding related applications in many scenarios, such as autonomous driving and robotics. However, the deep learning of a 3D point cloud still faces major challenges, such as small dataset size. **Method** In this study, the depth map sequence is first converted into a 3D point cloud sequence to represent human behavior information, and the large and authoritative datasets in the depth dataset are converted into point cloud datasets to compensate for the shortcoming of the small size of point cloud datasets. Given the huge amount of point cloud data, the traditional point cloud deep learning network will use a sampling algorithm to sample the point cloud before feature extraction. The most commonly used algorithm is random subsampling, which will inevitably lead to the destruction of point cloud structural information. To improve the utilization rate of temporal and spatial structure information and compensate for the loss of such information during the random subsampling of a point cloud, a point cloud human behavior recognition network that combines coordinate transformation and spatiotemporal information injection is proposed for motion recognition in this study. The network consists of two modules: the feature extraction module and the spatiotemporal information injection module. The feature extraction module extracts the deep appearance contour features of the point cloud through operations, such as the abstraction manipulation layer, multilayer perceptron, and maximum pooling. Among which, the abstraction manipulation layer includes the sampling, grouping, convolutional block attention module (CBAM), and PointNet layers. In the spatiotemporal information injection module, time sequence and spatial structure information are injected for abstract features. When timing information is injected, the sine and cosine functions of different frequencies are used as time position coding, because sine and cosine functions are unique and robust in the position of each vector in the disordered direction. During spatial structure information injection, the abstract features after location coding are multiplied with a group of learnable normal distribution random tensors and projected onto the corresponding dimension space. Then, the coefficients of the random tensors are learned through the network to find the optimal projection space that can better focus on the structural relations between point clouds. Subsequently, the feature enters the interpoint attention mechanism module to further learn the structural relationship between point cloud data points and points through the interpoint attention mechanism. Finally, the multilevel features in feature extraction and information injection are aggregated and inputted into the classifier for classification. **Result** A large number of experiments are performed on three common datasets, and the proposed network structure exhibits good performance. Accuracy on the NTU RGB+d60 datasets is 1. 3% and 0. 2% higher than those of PSTNet and SequentialPointNet, respectively, considerably exceeding the recognition accuracy of other networks. Although the accuracy of the NTU RGB+d120 dataset is 0. 1% lower than that of SequentialPointNet, it remains in a leading position compared with other networks. The network recognition accuracy proposed in this study is 1. 9% higher than that of PSTNet. The NTU dataset is one of the largest human action datasets. To ensure the robustness of the network model, the effect of the point cloud human behavior recognition network that combines coordinate transformation and spatiotemporal information injection on small datasets is verified, and experimental comparison was performed on small datasets of MSR Action3D. The recognition accuracy of the network proposed in this study was 1. 07% higher than that of SequentialPointNet, and considerably higher than those of other networks. **Conclusion** In this study, we propose a point cloud human behavior recognition network that combines coordinate transformation and spatiotemporal information injection for behavior recognition. Through coordinate transformation, the depth map sequence is converted into 3D point cloud sequence for the characterization of human behavior information, compensating for the shortcomings of insufficient depth information, spatial information, and geometric features, and improving the utilization rate of spatiotemporal structure information. The network proposed in this study not only obtains static point cloud contour features, but also integrates dynamic temporal and spatial information to compensate for the temporal and spatial losses caused by sampling during feature extraction.

Key words: human behavior recognition; coordinate transformation; point cloud sequence; feature extraction; spatiotemporal information

0 引 言

随着计算机视觉的不断发展,行为识别在视频监控和人机交互等诸多领域中展现出广泛的应用前景和研究价值。利用深度图序列(许艳 等,2018;李兴 等,2019;施海勇 等,2023)进行人体行为识别是机器视觉和人工智能中的一个重要研究领域,广泛使用的深度图序列尽管可以提供深度信息,但易受其他因素影响,行为数据的时空结构信息大量丧失。点云(Guo等,2021b;陶帅兵 等,2021)的出现弥补了深度图数据的劣势。点云就是分布在三维空间中的离散点集,它对复杂场景以及物体的外形表达具有独特的优势,但由于点云分布不规则且无序的性质,在点云上应用深度学习是不容易的。点云学习可分为基于多视图的、基于体积的和基于点的方法。基于多视图的方法首先将一个三维形状投影到多个视图中,并提取视图特征,然后融合这些特征进行精确的形状分类;基于体积的方法通常是将点云体素化为三维网格,然后应用三维卷积神经网络(convolutional neural network, CNN)对其进行形状分类;基于点的方法根据每个点的特征学习所使用的网络架构,独立地对每个点建模,然后使用对称聚合函数聚合全局特征。PointNet(Qi等,2017a)是点云深度学习的开山之作。PointNet的核心思想是利用一组多层感知机(multilayer perceptron, MLP)抽象每个点来学习其对应的空间编码,然后通过一个对称函数将所有单独的点特征集合起来得到一个全局的点云特征。但是PointNet缺乏对局部特征的提取及处理,而且现实场景中的点云往往是疏密不同的,而PointNet是基于均匀采样的点云进行训练的,导致了其在实际场景中准确率的下降。因此提出了一个分层网络PointNet++(Qi等,2017b),PointNet++的特征提取由3部分组成,分别为采样层、分组层和PointNet层,这3个层构成一个抽象层,PointNet++由几个抽象操作集合组成,PointNet++通过几个抽象层的层级结构逐步利用局部区域信息学习特征,网络结构更具有鲁棒性,但随机的最远距离点采样(farthest point sample, FPS)不可避免地会损失点云数据的时空信息。

为了解决上述问题,本文提出了一种结合坐标转换和时空信息注入的点云人体行为识别网络,该网络将深度图序列进行了信息转换,生成点云序列,并对其进行时空建模。网络由两个模块组成,即特征提取模块和时空信息注入模块。特征提取模块将每个点云框架抽象为一个外观轮廓的特征向量,以此来捕捉复杂的时空结构。在时空信息注入模块中,为点云的外观轮廓特征向量注入时空信息,其中借助可学习的正态分布随机张量的方法寻找空间结构信息上的特征变化,不仅能更好地表示数据的空间结构信息,也能加快网络的运行速度。在进行三维动作识别之前,将网络中的不同尺度特征串联起来。在结合坐标转换和时空信息注入的点云人体行为识别网络中,不同的点云框架在最终的分类网络层之前共享相同的网络架构和网络权重。

本文的主要贡献如下:1)提出一种结合坐标转换和时空信息注入的点云人体行为识别网络,通过点云特征提取模块和时空信息注入模块,解决了深度图序列时空结构信息的利用率不足的问题;2)通过构造时空信息注入模块,为静态点云序列注入动态信息(点云序列间的时序信息和运动帧的空间结构信息),弥补了点云抽象操作下采样时部分信息丢失的不足;3)设计了点间注意力机制模块,通过可学习的正态分布随机张量将数据映射到相应的空间中,不断寻找最优的投影空间,得到最佳的空间结构信息权重矩阵,以此表征运动帧的空间结构特征。用运动帧的空间结构特征替代点云帧的点特征。

1 相关工作

由于点云分布不规则且无序的性质,在点云上应用深度学习是不容易的,基于点云序列的三维人体动作识别是一项具有挑战性的新任务。PointNet是点云深度学习的开创之举。PointNet利用多层感知机、最大池化和刚性变化来保证置换和旋转下的不变性。PointNet++在此基础上通过几个抽象层的层级结构逐步学习局部特征,网络结构更具有鲁棒性。点云数据在时空维度上展现了不规则性和无序

性，不同帧中点的出现也无法保证一致性。为此Fan 等人（2022）提出了 PST 卷积（point spatio-temporal convolution）来编码点云序列的时空局部结构。PST 卷积首先解开点云序列的时空纠缠。此外，将 PST 卷积用分层的方式合并到一个深网络 PSTNet 中模拟点云序列。为了避免点跟踪，Fan 等人（2021）提出了 P4Transformer（point 4D Transformer）网络建模点云视频。P4Transformer 包括一个点 4D 卷积和一个 Transformer。Xu 等人（2021）介绍了一种用于三维点云处理的通用卷积运算 PAConv（position adaptive convolution），通过动态组装存储在权重库中的基本权重矩阵来构造卷积核，使得 PAConv 比 2D 卷积具有更大的灵活性，可以更好地处理不规则且无序的点云数据。Li 等人（2023）对称构造了两个点云特征图，从点云序列中识别人类行为，即点云外观图（point cloud appearance map，PCAM）和点云运动图（point cloud motion map，PCMM）。为了构建 PCAM，Li 等人（2023）设计了一种类似 MLP 的网络架构，用于在虚拟动作序列中捕获人类动作的时空外观特征；使用类似 MLP 的网络架构在虚拟动作差分序列中捕获人体动作的运动特征来构建 PCMM，最后，将两个点云特征图描述符连接起来并发送到一个全连接的分类器，以进行人类行为识别。

此外，Transformer 也逐渐应用于图像视觉任务，且效果优于流行的卷积网络。其中，Guo 等人（2021a）提出了一种新的点云学习框架 PCT（point cloud Transformer），PCT 的核心思想是利用 Transformer 固有的顺序不变性，避免定义点云数据的顺序，并通过注意力机制进行特征学习，注意力权重的分布与部分语义高度相关，并且不会随空间距离而严重衰减。Song 等人（2022b）提出了一种用于三维点云分析的新型增强型局部语义学习 Transformer，其中局部语义学习点云互感器（local semantic learning point cloud Transformer，LSLPCT）不仅可以学习 3D 点云的全局信息，还可以端到端地增强对局部语义信息的感知，局部语义学习自我注意机制（local semantic learning self-attention，LSL-SA）可以并行感知全局上下文信息并捕获更细粒度的局部语义特征。Liu 等人（2022）提出了一个新的端到端优化双流框架，称为几何 Transformer（geometrymotion-Transformer，GMT），GMT 使用特征提取模块（feature extraction module，FEM）在不使用体素化过程的情况下在帧之间生成一对一的对应关系，从原始点云中显式提取几何和多尺度运动表示，并提出了一种改进的基于 Transformer 的特征融合模块（feature fusion module，FFM），以有效地融合双流特征。

结合坐标转换和时空信息注入的点云人体行为识别网络根据将点云的时间和空间维度进行解耦，处理每个点云框架的空间结构和时间变化，从而进行时空特征提取。使用位置编码为点云抽象特征加入时序信息，通过可学习的随机张量对空间结构进行投影，寻找最佳的空间结构信息权重。最后将网络中不同层次的特征聚合后进行行为识别。

2 网络结构介绍

本文提出的结合坐标转换和时空信息注入的点云人体行为识别网络总体结构如图 1 所示。网络由特征提取模块和时空信息注入模块组成，在特征提取模块中，输入每一帧的点云集，输出对应帧外观轮廓的时空特征向量，以此表征时空信息。通过时空信息注入模块给所有帧加入时序信息和空间尺度信息。之后将多尺度的人体运动特征数据和时空特征数据有效融合，并利用全连接神经网络进行动作分类识别。

2.1 深度坐标系到点云坐标系的转换

人体行为识别的研究大量采用了深度图像序列。与 RGB 图像相比，深度图像基本不受自然光线影响，并提供了三维信息数据，但该数据只代表在可视范围内目标与深度摄像机的距离，数据冗余量大，对时空结构信息的表达也不充分。点云是在同一空间参考系下表达目标空间分布和目标表面特性的海量的点集合。点云的获取方式有多种，如通过各种类型的 3D 扫描仪、激光雷达和 RGB-D 相机。点云数据可以提供丰富的几何、形状和尺度信息，这是深度图所不能比拟的。通过坐标转换将深度图序列转换为点云序列，可以很容易地找到相邻点信息，弥补了深度图数据的不足。

深度图到点云数据的转换通常采用坐标系变换的方法，通过将图像坐标系转换为世界坐标系，深度图转换为点云数据。其中，图像坐标系转换为世界坐标系计算为

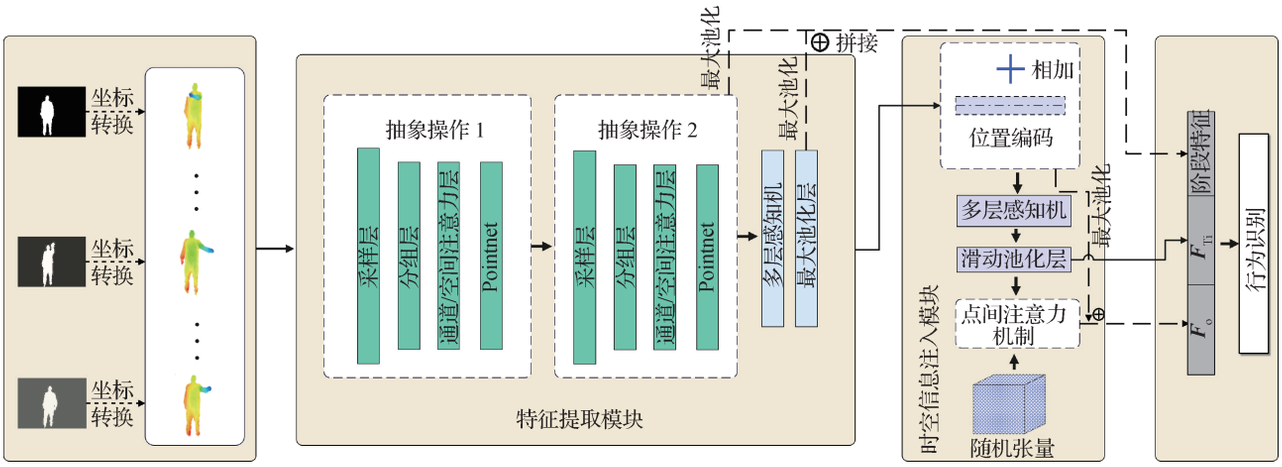

图1　结合坐标转换和时空信息注入的点云人体行为识别网络模块图

Fig. 1　Module diagram of human behavior recognition network in point cloud based on coordinate transformation and spatiotemporal information injection

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = D \begin{bmatrix} \frac{1}{f_x} & 0 & 0 \\ 0 & \frac{1}{f_y} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \tag{1}$$

式中，x，y，z为点云坐标系，D为深度值，f_x，f_y分别为镜头x，y方向的焦距，x'和y'是图像坐标系。得到图像点到世界坐标点的变换关系，具体为

$$x = \frac{\left(x' - c_x\right) \times D}{f_x} \tag{2}$$

$$y = \frac{\left(y' - c_y\right) \times D}{f_y} \tag{3}$$

$$z = D \tag{4}$$

式中，c_x，c_y分别是光心在图像坐标系下的坐标。

通过上述公式的变化，深度图序列中的每一帧深度图像转换成对应的点云帧，组成点云序列，相应深度数据集转换为点云数据集后作为网络的输入，如图2所示。

2.2　特征提取模块

受PointNet++的启发，本文构建了特征提取模块。该模块由两个抽象操作层、一组多层感知机和最大池化层组成。

抽象操作层由采样层、分组层、通道，空间注意力层（convolutional block attention module，CBAM）和PointNet层组成。

1）在采样层，使用最远距离点采样（FPS）从N个点的点集中选择n个点，降低数据集规模。FPS算法的流程为：首先随机选取一个点作为初始点加入初始点集，计算剩余点到初始点的欧氏距离，选距离最远的点加入到初始点集中，然后计算其余点到初始点集的距离，其余点中某个点到初始点集中所有点的欧氏距离中最小的值作为这个点到初始点集的距离，选取其余点中到初始点集距离最大的点加入初始点集，以此类推，直到初始点集长度为n。寻找初始点集及FPS算法的过程描述为

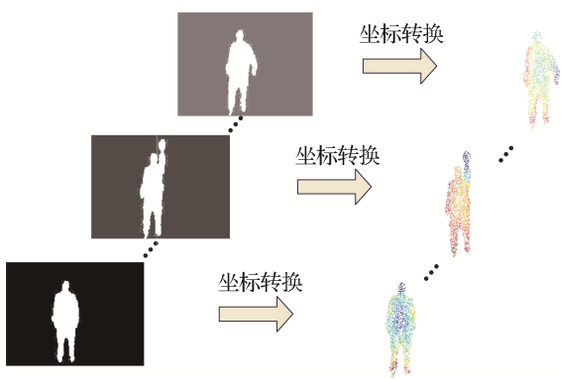

图2　深度序列转换为点云序列

Fig. 2　Graph of depth sequence to point cloud sequence

$$\max \left\| \boldsymbol{x}_i - \boldsymbol{P} \right\| \tag{5}$$

$$\text{s.t.} \begin{cases} \boldsymbol{P} = \left\{ \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{i-1} \right\} \\ \left\| \boldsymbol{x}_j - \boldsymbol{P} \right\| = \min\left(\left\| \boldsymbol{x}_j - \boldsymbol{x}_1 \right\|, \cdots, \left\| \boldsymbol{x}_j - \boldsymbol{x}_{i-1} \right\| \right) \\ i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, N - i + 1 \end{cases}$$

式中，$\boldsymbol{P}$代表初始点集，$\| \boldsymbol{x} - \boldsymbol{P} \|$代表点到初始点集的欧氏距离，$\boldsymbol{x}_i$代表初始点集中以及即将加入初始点集的点，范围是1到n。$\boldsymbol{x}_j$代表初始点集外的其余点，范围为1到$N - i + 1$。定义$\boldsymbol{P}_t = \left\{ \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \right\}$

为第t帧的点云集，$\boldsymbol{P}_T=\{\boldsymbol{P}_t\}_{t=1}^{T}$为$T$帧的点云序列。

2)在分组层，通过质心点与周围相同半径内的局部点组成局部邻域，便于网络学习点与点之间的空间结构关系。球半径查询方法可以查找在质心点半径范围内所有点。第1个分组层的输入是一组大小为$n\times(d+c)$(具有d维坐标和c维点特征的n个点)的点集和一组大小为$n'\times d$的质心的坐标，输出是一组大小为$n_1\times k\times(d+c_1)$的点集，其中每组对应一个局部区域，k是质心点邻域中的点数。

3)在通道注意力和空间注意力层，使用通道注意力和空间注意力沿着通道和空间两个维度进行注意力权重学习，对点云特征进行自适应调整，获取重要特征，压缩不重要特征，表征每一帧人体行为静态外观的时间信息和空间结构，如图3所示。为了有效计算通道注意力，需要对输入特征图的空间维度进行压缩，对于空间信息的聚合，常用的方法是平均池化。另外，最大池化可以收集到难区分物体之间更重要的线索，以获得更详细的通道注意力，所以平均池化和最大池化的特征是同时使用的。因此，通道注意力模块同时使用平均池化和最大池化后的点云特征，然后将它们依次送入一个共享权重的多层感知机中，最后将输出的特征向量进行合并。空间注意力主要聚焦于哪部分的有效信息较丰富，这是对通道注意力的补充。通过最大池化和平均池化各获得一张特征图，而后将它们拼接成一张2D特征图，再送入标准7×7卷积进行参数学习，最终得到一幅1D的权重特征图，该图编码了需要关注的位置。从空间的角度来看，通道注意力是全局的，而空间注意力是局部的。本文CBAM模块的结构表达为

$$A(in)=\sigma\left(MLP\left(m(in)\right)+MLP\left(n(in)\right)\right) \tag{6}$$

式中，$A()$表示通道注意力和空间注意力操作，in表示模块的输入，MLP表示多层感知机操作，m和n表示平均池化和最大池化操作，σ表示激活函数。

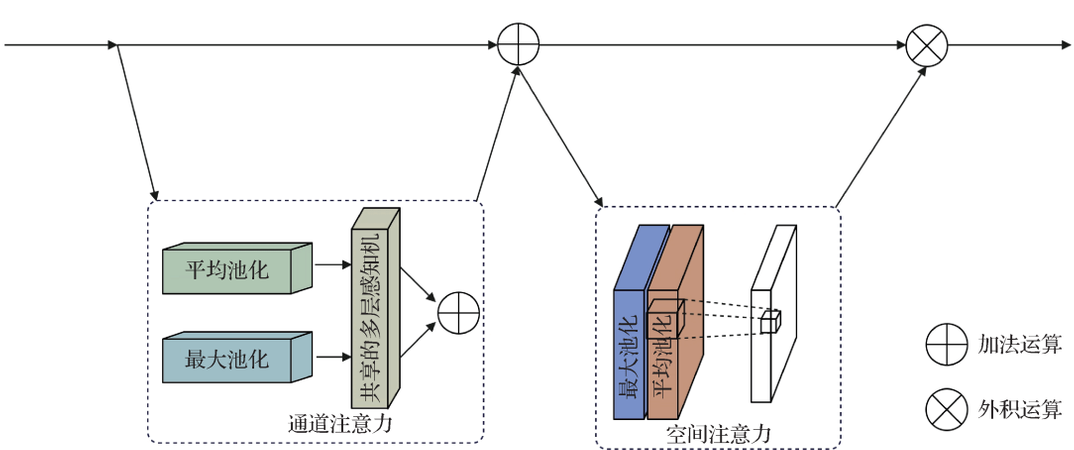

图3　通道注意力和空间注意力

Fig. 3　Channel attention and spatial attention

4)在PointNet层，由一组MLP和一个最大池化操作组成，通过MLP和最大池化操作来表征局部区域特征。在这一层中，输入的是数据为$n_1\times k\times(d+c_1+1)$的$n_1$个局部区域，输出数据为$n_1\times(d+c_1)$，由$n_1$个具有$d$维坐标的子采样点和总结本地上下文的新$c_1$维特征向量组成。输出中的每个局部区域都是其质心和质心邻域的局部抽象特征的连接。

抽象操作2与抽象操作1类似，输入的数据为$n_1\times(d+c_1)$，输出为$n_2\times(d+c_2)$，将输出记为$\boldsymbol{f}_{ab}$。

最后，通过一组多层感知机和最大池化层表征整个点云框架的时空信息，计算为

$$\boldsymbol{f}=MAX\left\{MLP(\boldsymbol{f}_{ab})\right\} \tag{7}$$

式中，$\boldsymbol{f}$为一帧点云帧通过多层感知机和最大池化操作后的特征向量，MAX表示最大池化操作，所有点云帧通过特征提取模块的输出为$\boldsymbol{F}=\{\boldsymbol{f}_t\}_{t=1}^{T}$，$T$为一个行为动作的总帧数，$\boldsymbol{f}$的大小为$1\times d_o$，$\boldsymbol{F}$的大小为$T\times d_o$，$d_o$为输出通道的大小。

2.3　时空信息注入模块

通过点云对深度图像进行信息表征弥补了深度图数据时空信息不足的缺点，但点云序列的转换以

及随机最远点采样会使原本的时空结构信息损失完整性，在一定程度上损失一部分时空结构信息，所以有必要对点云序列进行额外时空结构信息注入。

2.3.1 时序信息注入

由图1所示，经过特征提取模块形成的外观轮廓的时空特征向量序列 $\boldsymbol{F}=\{\boldsymbol{f}_t\}_{t=1}^{T}$ 在进入时空信息注入模块后首先进行时序信息注入。为了对人体动作的时间信息进行编码，使用位置编码、共享MLP层和滑动块最大池化层。位置编码层为特征向量序列注入时间位置信息。共享的MLP层对每个独立的特征向量执行一组MLP，以提取每个点云框架的时空信息。采用滑动块最大池化层在多个时间尺度上提取序列空间信息。

1）位置编码层。给定输入特征向量序列 $\boldsymbol{F}=\{\boldsymbol{f}_t\}_{t=1}^{T}$，通过加入位置编码注入顺序信息。因为正弦和余弦函数在无序方向中，每个向量的位置具有唯一性和很好的鲁棒性，所以使用不同频率的正弦和余弦函数作为时间位置编码。

$$\boldsymbol{PE}_{p,2l}=\sin\left(p/10\,000^{2l/d_m}\right) \tag{8}$$

$$\boldsymbol{PE}_{p,2l+1}=\cos\left(p/10\,000^{2l/d_m}\right) \tag{9}$$

式中，$\boldsymbol{PE}$ 表示二维矩阵，大小和 $\boldsymbol{f}_t$ 相同，p 表示时间位置。l 表示特征向量的位置，d_m 表示特征向量的维度。偶数位置使用正弦函数，奇数位置使用余弦函数。将位置编码函数与 $\boldsymbol{f}_t$ 聚合以此加入时间位置信息生成特征向量 $\hat{\boldsymbol{f}}_t$。$\hat{\boldsymbol{f}}_t$ 是经过位置编码后的新的特征向量。

2）共享的MLP层。经过时间位置嵌入层后，将顺序信息简单地嵌入到空间信息序列中。为了进一步提取时空信息，对每个特征向量应用一组MLP，即

$$\tilde{\boldsymbol{f}}_t=MLP\left(\hat{\boldsymbol{f}}_t\right) \tag{10}$$

式中，$\tilde{\boldsymbol{f}}_t$ 表示使用MLP操作更新的特征向量。

3）滑动块最大池化层。在这一层中，使用最大池化操作对多个特征向量进行聚合。为了捕获点云序列内的子动作和更有鉴别性的运动信息，提出滑动块最大池化策略，将向量序列 $\tilde{\boldsymbol{F}}=\{\tilde{\boldsymbol{f}}_t\}_{t=1}^{T}$ 分成与点云帧等量的块，其中前 e 个块组成滑动块，然后对滑动块进行最大池化操作，生成相应的子特征。之后将滑动块向后滑动 m 个点云帧距离，再进行最大池化操作并生成子特征，直到滑动块到达序列末为止。最后，所有的子特征被简单地连接起来，形成人类行为的时间子特征 $\boldsymbol{F}_{\mathrm{Ti}}$。

为了获得更充足的人体运动时空信息，从位置编码前的不同阶段整合人体动作特征（如图1中阶段特征），以此丰富时间特征序列。整合方法为

$$\boldsymbol{T}_{\mathrm{Time}}=\boldsymbol{F}_{\mathrm{Ti}}+MAX\left[MAX\left(\boldsymbol{F}_{\mathrm{ab}}\right)\right]+MAX\left(\boldsymbol{F}\right) \tag{11}$$

式中，$\boldsymbol{F}_{\mathrm{ab}}=\{\boldsymbol{f}_{\mathrm{ab}}\}_{t=1}^{T}$，$\boldsymbol{T}_{\mathrm{Time}}$ 为时间特征向量序列。

2.3.2 空间信息注入

Li等人（2022）指出了强空间结构和弱时间变化的人类行为特性，即当人们观察多帧的人体动作时，即使时间顺序杂乱，也可以通过静态外观表象进行大致有效的动作识别，说明空间结构信息表征在动作识别时的重要性，意味着点云序列动作识别中强空间结构信息的学习和表征对网络性能有着不可或缺的作用，而原始PointNet++中的抽象操作使用FPS采样，在加大感受野的同时，也不可避免地损失其余的空间信息。在经过滑动池化层后，将带有时序信息的特征向量称为三维向量关系序列（即 $\boldsymbol{F}_{\mathrm{Ti}}$）。如图4所示，三维向量关系序列同一组可学习的kaiming正态分布的随机张量进行乘积，将三维向量关系序列投影到相应的维度空间中，再通过网络学习随机张量的系数，寻找更能关注点云间结构关系的最优投影空间。

聚类之后进入点间注意力机制模块，通过点间注意力机制进一步学习点云数据点与点之间的结构关系，并生成可以表征点云数据空间结构关系的权重系数矩阵。

1）随机张量。为了更好地进行点云深度学习，让网络自主地学习到更适合表征数据空间结构的关系矩阵，采用一组设定好大小但数据随机的张量集，通过迭代不断学习更优的数据参数，寻找最优投影空间。张量是一种强大的表示方向和空间的方法，通过张量不仅能更好地表示数据的空间结构信息，也能加快网络的运行速度。

2）点间注意力机制。点间注意力机制由一组多层感知机和softmax函数等组成，多层感知机可以很好地学习到点云数据中更关键点的时空信息，再经过softmax函数层转换成权重系数，即生成了可以表征点云数据空间结构关系的权重系数矩阵，其表现形式为

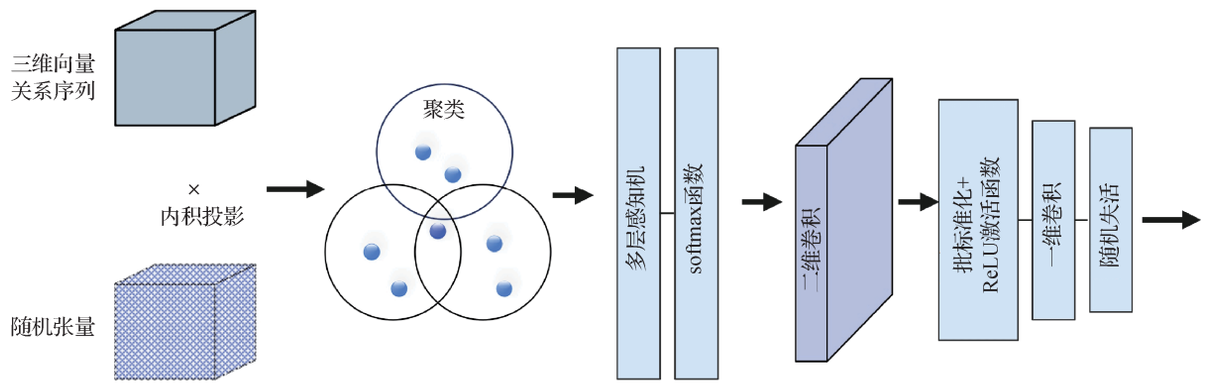

图 4　点间注意力机制（空间信息注入）

Fig. 4　Inter-point attention mechanism (spatial information injection)

$$\boldsymbol{F}_{\mathrm{s}}=\Phi\left(MAX\left\{MLP\left(C\left(\boldsymbol{R}\times\boldsymbol{F}_{\mathrm{Ti}}\right)\right)\right\}\right) \tag{12}$$

式中，$\boldsymbol{F}_{\mathrm{s}}$ 表示生成的可以表征点云数据空间结构关系的权重系数矩阵（时空特征 1），$\boldsymbol{R}$ 表示随机张量，C 表示聚类操作，Φ 表示特征映射操作，即为 softmax 后的卷积和批正则化等操作。

为了将点间关系与点云序列数据各点相结合，使用的方法为

$$\boldsymbol{F}_{\mathrm{o}}=MAX\left\{\boldsymbol{F}_{\mathrm{Ti}}\right\}\oplus\boldsymbol{F}_{\mathrm{s}} \tag{13}$$

式中，$\boldsymbol{F}_{\mathrm{Ti}}$ 为经过时序信息注入后生成的三维向量关系序列，将其抽象（时空特征 2）并与时空特征 1 结合，生成空间结构信息特征向量序列 $\boldsymbol{F}_{\mathrm{o}}$。

最后，将时间特征向量序列 $\boldsymbol{T}_{\mathrm{Time}}$ 和空间结构特征向量序列 $\boldsymbol{F}_{\mathrm{o}}$ 进行简单的拼接，然后发送到一组全连接层中进行人类动作识别。

3　实　验

3.1　数据集

在两个大型公共动作识别数据集 NTU RGB+d60（Shahroudy 等，2016）和 NTU RGB+d120（Liu 等，2020a）以及一个小型公共数据集 MSR Action3D（Li 等，2010）上评估了所提出的方法。

NTU RGB+d60 数据集由 60 个动作的 56 880 个深度视频序列组成，是最大的人类动作数据集之一。

NTU RGB+d120 数据集是目前最大的三维动作识别数据集，是 NTU RGBD 60 数据集的扩展。NTU RGB+d120 数据集由 120 个动作的 114 480 个深度视频序列组成。

MSR Action3D 数据集包含来自 10 个受试者的 20 个动作的 557 个深度视频样本，每个动作由每个受试者执行 2 或 3 次。

3.2　实现细节

首先，从点云集合中随机抽取 2 048 个点。然后，利用 PFS 算法从 2 048 个点中选取 512 个点。在特征提取模块中，对每个点云框架进行两次集合抽象操作，采用 SequentialPointNet 中获取的最佳参数设置。在第 1 组抽象操作中，选择 128 个质心来确定点组，组半径设置为 0. 06。每个点组中的点数设置为 48。在第 2 组抽象操作中，选择 32 个质心来确定点组，组半径设置为 0. 1。每个点组的点数设置为 16，如表 1 所示。在进行提取空间结构信息前，首先使用聚类生成三维向量关系序列，聚类半径设置为 20。在进行提取空间结构信息时，随机张量大小设置为（8，64，64），dropout 设置为 0. 5。用 Adam 作为优化器。学习速率从 0. 001 开始，每 10 个 epoch 以 0. 5 的速率衰减，使用交叉熵损失函数。

表 1　特征提取实验设置

Table 1　Feature extraction experiment set

抽象操作	质心点个数	组半径	点组个数
第 1 组	128	0.06	48
第 2 组	32	0.10	16

3.3　实验过程

为了探索哪种数据更有利于空间信息的提取，以及不同数据库对于不同数据提取方式的效果，本文进行了不同的对比实验，寻找最适合的实验方法。

使用 MSR Action3D 小数据集进行实验，首先使用两种不同的数据作为时空信息注入模块的输入，其中之一为原始三维点云数据，即为抽象操作之前的三维点云数据；另外一种数据为经过位置编码，已

经进行特征提取后，通过聚类生成的三维向量关系序列（以下分别称为原始数据和关系数据）。之后进行多次实验并记录最后的实验结果，如表2所示。

由表2实验1—实验4可以看出，当批次大小相同都设置为8、迭代次数为100时，使用原始数据作为输入且注入时空特征的准确率为89.71%，使用关系数据作为输入且注入时空特征的准确率为91.91%，而当批次大小设置为150时，使用原始数据的准确率为92.65%，使用关系数据的准确率达到了93.01%。由此可见，使用关系数据作为输入比使用原始数据作为输入效果更优。再结合实验6和7可得出结论，当迭代次数为150时，准确率趋于平稳且最优。

由表2实验4—实验9可以看出，当批次大小都设置为8、迭代次数为150时，使用关系数据作为输入的前提下，只注入时空特征1或时空特征2的准确率分别为86.76%和91.18%，均低于未注入时空特征的准确率，其中只注入时空特征1的准确率比原来低5.18%，而将时空特征1与时空特征2融合后注入，准确率达到93.01%。由此可见注入完整时空特征的重要性。再由表2中实验4和5可知，MSR Action3D小数据集上的批次大小设置为8最为合适。

使用MSR Action3D小数据集得出结果后，将参数迁移，开始对NTU RGB+d120和NTU RGB+d60大数据集进行实验，使用关系数据作为时空信息注入模块的输入，并记录结果，如表3所示。

表2　MSR Action3D数据集上的实验过程

Table 2　The experimental process on MSR Action3D dataset

实验序号	输入	批次大小	迭代次数	准确率/%	备注
1	原始数据	8	100	89.71	注入时空特征
2	原始数据	8	150	92.65	注入时空特征
3	关系数据	8	100	91.91	注入时空特征
4	关系数据	8	150	93.01	注入时空特征
5	关系数据	16	150	90.81	注入时空特征
6	关系数据	8	100	91.58	未注入时空特征
7	关系数据	8	150	91.94	未注入时空特征
8	关系数据	8	150	86.76	注入时空特征1
9	关系数据	8	150	91.18	注入时空特征2

表3　NTU RGB+d60/120数据集上的实验过程

Table 3　The experimental procedure on NTU RGB+d60/120 dataset

实验序号	数据集	输入	批次大小	迭代次数	准确率/%	备注
1	NTU RGB+d60	原始数据	16	150	97.59	注入时空特征
2	NTU RGB+d60	原始数据	32	150	97.82	注入时空特征
3	NTU RGB+d60	关系数据	64	150	97.67	注入时空特征
4	NTU RGB+d120	关系数据	32	150	93.90	注入时空特征
5	NTU RGB+d120	关系数据	48	150	95.34	注入时空特征
6	NTU RGB+d120	关系数据	64	150	94.42	注入时空特征

通过实验对比寻找NTU RGB+d60/120数据集最适合的批次大小。由表3实验1—实验3结果可知，准确率的大小与批次大小不是正相关关系，当批次大小设置为32时，结果为97.82%且最优，当批次大小为16和64时，准确率有所下降。在NTU RGB+d120大数据集上，准确率的大小与批次大小也不是正相关的关系，当批次大小设置为48时，结果为95.34%且最优，这也直接证明了时空信息注入的合

理性和可行性。由NTU数据集的实验可得出结论，该网络模型结构对于人体行为识别的分类具有较好的优越性。

3.4 与最先进的方法比较

为了验证网络的性能，在NTU RGB+d60数据集、NTU RGB+d120数据集和MSR Action3D数据集上实现了与其他先进方法的对比实验。

1）NTU RGB+d60数据集。首先比较结合坐标转换和时空信息注入的点云人体行为识别网络和NTU RGB+d60数据集上的最先进的方法。NTU RGB+d60数据集是一种大规模的室内人类活动数据集。如表4所示，结合坐标转换和时空信息注入的点云人体行为识别网络的准确率达到了97.8%。本文方法表现出与其他方法相当甚至更好的性能，达到了最先进的性能。

2）NTU RGB+d120数据集。将结合坐标转换和时空信息注入的点云人体行为识别网络与NTU RGB+d120数据集上的最先进的方法进行比较。NTU RGB+d120数据集是用于3D动作识别的最大数据集。与NTU RGB+d60数据集相比，在NTU RGB+d120数据集上进行三维人体动作识别更具挑战性。如表5所示，结合坐标转换和时空信息注入的点云人体行为识别网络的准确率达到了95.3%，仅低于SequentialPointNet，并且展现出比其他网络更优秀的性能。

3）MSR Action3D数据集。为了综合评价本文方法，在小型MSR Action3D数据集上进行了对比实验。为了缓解小尺度数据集上的过拟合问题，将批量大小设置为8，其他参数设置与两个大规模数据集上的设置相同。表6展示了不同方法的识别精度，结合坐标转换和时空信息注入的点云人体行为识别网络在MSR Action3D数据集上取得了最先进的性能。

表4 NTU RGB+d60数据集上的行为识别准确率

Table 4 Behavior recognition accuracy on NTU RGB+d60 dataset

方法	输入	识别率/%
MVDI(2019)（Xiao等，2019）	深度图	87.3
3DFCNN(2020)（Sánchez-Caballero等，2022）	深度图	80.4
ConvLSTM(2020)（Sanchez-Caballero等，2020）	深度图	79.9
ST-GCN(2018)（Yan等，2018）	骨骼图	88.3
DGNN(2019)（Shi等，2019）	骨骼图	96.1
DDGCN(2020)（Korban和Li，2020）	骨骼图	97.1
3s-CrosSCLR(2021)（Li等，2021a）	骨骼图	92.5
Sym-GNN(2021)（Li等，2021b）	骨骼图	96.4
LST(2022)（Xiang等，2023）	骨骼图	97.0
Info-GCN(2022)（Chi等，2022）	骨骼图	96.9
EfficientGCN-B4(2022)（Song等，2022a）	骨骼图	95.7
3DV-PointNet++(2020)（Wang等，2020）	点	96.3
P4Transformer(2021)	点	96.4
PSTNet(2021)	点	96.5
SequentialPointNet(2021)	点	97.6
本文	点	**97.8**

注：加粗字体表示最优结果。

表5 NTU RGB+d120数据集上的行为识别准确率

Table 5 Behavior recognition accuracy on NTU RGB+d120 dataset

方法	输入	识别率/%
Baseline(2020)（Liu等，2020a）	深度图	40.1
ST-GCN(2018)	骨骼图	88.3
MS-G3D Net(2020)（Liu等，2020b）	骨骼图	88.4
4s Shift-GCN(2020)（Cheng等，2020）	骨骼图	87.6
3s-crosSCLR(2021)	骨骼图	80.4
LST(2022)	骨骼图	91.1
Info-GCN(2022)	骨骼图	92.9
EfficientGCN-B4(2022)	骨骼图	89.1
3DV-PointNet++(2020)	点	93.5
P4Transformer(2021)	点	93.5
PSTNet(2021)	点	93.5
SequentialPointNet(2021)	点	**95.4**
本文	点	95.3

注：加粗字体表示最优结果。

根据表4—表6的对比结果可知，在NTU两个数据集上，本文方法领先于绝大部分网络，展现出较好的准确率优势，而在MSR Action3D小数据集上，本文方法以明显的优势领先于其他网络，其中准确率比SequentialPointNet提升了1.07%。由此可见，本文方法在大数据集和小数据集上都表现良好，尤其

表6 MSR Action3D数据集上的行为识别准确率
Table 6 Behavior recognition accuracy on MSR Action3D dataset

方法	输入	识别率/%
Kläser等人(2008)(Kläser等,2008)	深度图	81.43
Vieira等人(2012)(Vieira等,2012)	深度图	78.20
Actionlet(2012)(Wang等,2012)	骨骼图	88.21
MeteorNet(2019)(Liu等,2019)	点	88.50
PointNet++(2020)	点	61.61
P4Transformer(2021)	点	90.94
PSTNet(2021)	点	91.20
SequentialPointNet(2021)	点	91.94
本文	点	**93.01**

注:加粗字体表示最优结果。

更有利于小数据集的识别。

本文提出的结合坐标转换和时空信息注入的点云人体行为识别网络为了提高时空结构信息的利用率,提出特征提取模块和时空信息注入模块,为静态点云序列注入动态信息,弥补了点云的不足。其中点间注意力机制可以寻找最优的投影空间,得到了最佳的空间结构表征,这也导致了本文方法良好的性能。

为了进一步证明结合坐标转换和时空信息注入的点云人体行为识别网络的性能,在原来识别率指标的基础上引入NTU RGB+d60数据集和NTU RGB+d120数据集的另外3个指标cross-subject、cross-view和cross-setwp。不同指标的区别为训练集和测试集划分方式的不同。NTU RGB+d60和NTU RGB+d120的cross-subject根据受试者ID划分;NTU RGB+d60的cross-view根据相机ID划分;NTU RGB+d120的cross-steup指定id为偶数的样本进行训练,id为奇数的样本进行测试。实验结果如表7和表8所示。本文方法在8个结果中仅NTU RGB+d120上的cross-setup低于SequentialPointNet 0.1%。其中,在NTU RGB+d60上的cross-subject和cross-setup识别率分别高于SequentialPointNet 0.3%和0.2%,在NTU RGB+d120上的cross-subject识别率高于SequentialPointNet 0.5%,这也进一步表明了本文方法的优越性。

在SequentialPointNet的时空结构中,空间结构和时间变化是独立建模的,SequentialPointNet提出的强空间结构和弱时间变化的观念,SequentialPointNet着重强调对空间结构特征的提取。SequentialPointNet认为将空间信息和时间信息同等对待是不合理的,因为人的行为在空间维度上是复杂的,而在时间维度上是简单的。本文方法同等对待时间和空间特征的地位,在最终特征聚合阶段,时间特征和空间特征以同等维度大小融合。在某些动作,例如NTU RGB+d120中的嗅闻(A117)或耳语(A79)等微小动作(这类动作id大多为奇数)中,空间结构的重要性大于时序信息,这导致本文方法在NTU RGB+d120上的cross-setup识别率相比于SequentialPointNet较低。

表7 SequentialPointNet与本文方法在NTU RGB+d60数据集上的对比实验
Table 7 Comparison of SequentialPointNet and the method of ours on NTU RGB+d60 dataset

方法	cross-subject/%	cross-view/%
SequentialPointNet	90.3	97.6
本文	**90.6**	**97.8**

注:加粗字体表示最优结果。

表8 SequentialPointNet与本文方法在NTU RGB+d120数据集上的对比实验
Table 8 Comparison of SequentialPointNet and the method of ours on NTU RGB+d120 dataset

方法	cross-subject/%	cross-setup/%
SequentialPointNet	83.5	**95.4**
本文	**84.0**	95.3

注:加粗字体表示最优结果。

4 结 论

本文提出了一个结合坐标转换和时空信息注入的点云人体行为识别网络。该网络采取坐标转换的方式,将深度图序列转换为三维点云序列进行人体行为信息的表征,弥补了深度信息空间信息与几何特征不足的缺点,提高了时空结构信息的利用率。网络由两个模块组成,即特征提取模块和时空信息注入模块。特征提取模块提取点云序列的空间结构特征和时间变化特征。为了捕获时空结构,使用两个抽象操作将每个点云框架抽象为一个外观轮廓的

特征向量。在时空信息注入模块中，采用时间位置编码和滑动池化策略对特征向量序列进行时序信息注入。此外，通过一组可学习的正态分布随机张量寻找最优的投影空间，在最优投影空间中，通过点间注意力机制输出最佳的空间结构信息权重系数矩阵，为了保留原有的空间结构，系数矩阵与三维向量关系序列进行特征聚合，从而注入空间结构信息。最后对人体动作的多层次特征进行了融合与分类。在本文方法中，不同的点云框架共享相同的网络架构和权重。

在3个公共数据集上进行的大量实验表明，结合坐标转换和时空信息注入的点云人体行为识别网络展现了其优异的性能，其中，在MSR Action 3D数据集上，本文方法以明显的优势领先于其他网络，准确率比SequentialPointNet提升了1.07%；本文方法在NTU RGB+d120数据集上的准确率仅次于SequentialPointNet。原因在于SequentialPointNet与本文方法在时空特征权重的处理上不同。SequentialPointNet更加侧重于对空间结构特征的提取，对于微小动作的分类更加准确，因此，在cross-setup指标下本文方法的准确率比SequentialPointNet低0.1%。但在cross-subject和cross-view指标下，本文方法均比SequentialPointNet准确率高0.2%以上。

由于NTU数据集的规模较大，将训练小数据集的网络参数迁移，从而进行训练大数据集并不能完全展现网络的性能，下一步研究应探究不同的网络参数对于大数据集行为识别的影响，并增强网络的轻便性。未来工作将聚焦在研究点云人体行为识别的轻量性和实用性方面。在探究降低参数量实现网络轻量化的同时，设计适用于不同动作的时空特征融合方式，从而加强网络对不同动作，特别是微小动作的识别能力，提高网络的泛化性，并将结合坐标转换和时空信息注入的点云人体行为识别网络进一步应用于智能驾驶等领域中。

参考文献(References)

Cheng K, Zhang Y F, He X Y, Chen W H, Cheng J and Lu H Q. 2020. Skeleton-based action recognition with shift graph convolutional network//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 180-189 [DOI: 10.1109/CVPR42600.2020.00026]

Chi H G, Ha M H, Chi S, Lee S W, Huang Q X and Ramani K. 2022. InfoGCN: representation learning for human skeleton-based action recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20154-20164 [DOI: 10.1109/CVPR52688.2022.01955]

Fan H H, Yang Y and Kankanhalli M. 2021. Point 4D Transformer networks for spatio-temporal modeling in point cloud videos//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14199-14208 [DOI: 10.1109/CVPR46437.2021.01398]

Fan H H, Yu X, Ding Y H, Yang Y and Kankanhalli M S. 2022. PSTNet: point spatio-temporal convolution on point cloud sequences [EB/OL]. [2023-02-04]. https://arxiv.org/pdf/2205.13713.pdf

Guo M H, Cai J X, Liu Z N, Mu T J, Martin R R and Hu S M. 2021a. PCT: point cloud Transformer. Computational Visual Media, 7(2): 187-199 [DOI: 10.1007/s41095-021-0229-5]

Guo Y L, Wang H Y, Hu Q Y, Liu H, Liu L and Bennamoun M. 2021b. Deep learning for 3D point clouds: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12): 4338-4364 [DOI: 10.1109/TPAMI.2020.3005434]

Kläser A, Marszałek M and Schmid C. 2008. A spatio-temporal descriptor based on 3D-gradients//Proceedings of the British Machine Vision Conference. Leeds, UK: BMVC [DOI: 10.5244/C.22.99]

Korban M and Li X. 2020. DDGCN: a dynamic directed graph convolutional network for action recognition//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 761-776 [DOI: 10.1007/978-3-030-58565-5_45]

Li L G, Wang M S, Ni B B, Wang H, Yang J C and Zhang W J. 2021a. 3D human action representation learning via cross-view consistency pursuit//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 4739-4748 [DOI: 10.1109/CVPR46437.2021.00471]

Li M S, Chen S H, Chen X, Zhang Y, Wang Y F and Tian Q. 2021b. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 3316-3333 [DOI: 10.1109/TPAMI.2021.3053765]

Li W Q, Zhang Z Y and Liu Z C. 2010. Action recognition based on a bag of 3D points//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. San Francisco, USA: IEEE: 9-14 [DOI: 10.1109/CVPRW.2010.5543273]

Li X, Hou Z J, Liang J Z and Chang X Z. 2019. Bi-directional removal of reverse gravitational acceleration based on data segmentation. Journal of Computer-Aided Design and Computer Graphics, 31(4): 560-572 (李兴, 侯振杰, 梁久祯, 常兴治. 2019. 分段双向去除反向重力加速度算法. 计算机辅助设计与图形学学报, 31(4): 560-572) [DOI: 10.3724/SP.J.1089.2019.17344]

Li X, Huang Q, Wang Z J, Hou Z J and Yang T J. 2022. Sequential-PointNet: a strong parallelized point cloud sequence classification network for 3D action recognition [EB/OL]. [2023-02-04]. https://arxiv.org/pdf/2111.08492v1.pdf

Li X, Huang Q, Zhang Y F, Yang T J and Wang Z J. 2023. PointMap-Net: point cloud feature map network for 3D human action recognition. Symmetry, 15(2): #363 [DOI: 10.3390/sym15020363]

Liu J, Shahroudy A, Perez M, Wang G, Duan L Y and Kot A C. 2020a. NTU RGB+d120: a large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2684-2701 [DOI: 10.1109/TPAMI.2019.2916873]

Liu J H, Guo J Y and Xu D. 2022. GeometryMotion-Transformer: an end-to-end framework for 3D action recognition. IEEE Transactions on Multimedia, 25: 5649-5661 [DOI: 10.1109/TMM.2022.3198011]

Liu X Y, Yan M Y and Bohg J. 2019. MeteorNet: deep learning on dynamic 3D point cloud sequences//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (CVPR). Seoul, Korea (South): IEEE: 9245-9254 [DOI: 10.1109/ICCV.2019.00934]

Liu Z Y, Zhang H W, Chen Z H, Wang Z Y and Ouyang W L. 2020b. Disentangling and unifying graph convolutions for skeleton-based action recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 140-149 [DOI: 10.1109/CVPR42600.2020.00022]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5105-5114

Sánchez-Caballero A, de López-Diz S, Fuentes-Jimenez D, Losada-Gutiérrez C, Marrón-Romera M, Casillas-Pérez D and Sarker M I. 2022. 3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information. Multimedia Tools and Applications, 81(17): 24119-24143 [DOI: 10.1007/s11042-022-12091-z]

Sanchez-Caballero A, Fuentes-Jimenez D and Losada-Gutiérrez C. 2020. Exploiting the ConvLSTM: human action recognition using raw depth video-based recurrent neural networks [EB/OL]. [2023-02-04]. http://arxiv.org/pdf/2006.07744.pdf

Shahroudy A, Liu J, Ng T T and Wang G. 2016. NTU RGB+d: a large scale dataset for 3D human activity analysis//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 1010-1019 [DOI: 10.1109/CVPR.2016.115]

Shi H Y, Hou Z J, Chao X and Zhong Z K. 2023. Multimodal spatial-temporal feature representation and its application in action recognition. Journal of Image and Graphics, 28(4): 1041-1055 (施海勇, 侯振杰, 巢新, 钟卓锟. 2023. 多模态时空特征表示及其在行为识别中的应用. 中国图象图形学报, 28(4): 1041-1055) [DOI: 10.11834/jig.211217]

Shi L, Zhang Y F, Cheng J and Lu H Q. 2019. Skeleton-based action recognition with directed graph neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 7904-7913 [DOI: 10.1109/CVPR.2019.00810]

Song Y F, Zhang Z, Shan C F and Wang L. 2022a. Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2): 1474-1488 [DOI: 10.1109/TPAMI.2022.3157033]

Song Y P, He F Z, Duan Y S, Si T Z and Bai J W. 2022b. LSLPCT: an enhanced local semantic learning Transformer for 3-D point cloud analysis. IEEE Transactions on Geoscience and Remote Sensing, 60: #4708813 [DOI: 10.1109/TGRS.2022.3202823]

Tao S B, Liang C, Jiang T P, Yang Y J and Wang Y J. 2021. Sparse voxel pyramid neighborhood construction and classification of LiDAR point cloud. Journal of Image and Graphics, 26(11): 2703-2712 (陶帅兵, 梁冲, 蒋腾平, 杨玉娇, 王永君. 2021. 激光点云的稀疏体素金字塔邻域构建与分类. 中国图象图形学报, 26(11): 2703-2712) [DOI: 10.11834/jig.200262]

Vieira A W, Nascimento E R, Oliveira G L, Liu Z C and Campos M F M. 2012. STOP: space-time occupancy patterns for 3D action recognition from depth map sequences//Proceedings of the 17th Iberoamerican Congress. Buenos Aires, Argentina: Springer: 252-259 [DOI: 10.1007/978-3-642-33275-3_31]

Wang J, Liu Z C, Wu Y and Yuan J S. 2012. Mining actionlet ensemble for action recognition with depth cameras//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 1290-1297 [DOI: 10.1109/CVPR.2012.6247813]

Wang Y C, Xiao Y, Xiong F, Jiang W X, Cao Z G, Zhou J T and Yuan J S. 2020. 3DV: 3D dynamic voxel for action recognition in depth video//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 508-517 [DOI: 10.1109/CVPR42600.2020.00059]

Xiang W M, Li C, Zhou Y X, Wang B and Zhang L. 2023. Language action description prompts for skeleton-based action recognition. [EB/OL]. [2023-09-06]. http://arxiv.org/pdf/2208.05318.pdf

Xiao Y, Chen J, Wang Y C, Cao Z G, Zhou J T and Bai X. 2019. Action recognition for depth video using multi-view dynamic images. Information Sciences, 480: 287-304 [DOI: 10.1016/j.ins.2018.12.050]

Xu M T, Ding R Y, Zhao H S and Qi X J. 2021. PAConv: position adaptive convolution with dynamic kernel assembling on point clouds//Proceedings of 2021 IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR). Nashville, USA: IEEE: 3172-3181 [DOI: 10.1109/CVPR46437.2021.00319]

Xu Y, Hou Z J, Liang J Z, Chen C, Jia L and Song Y. 2018. Action recognition using weighted fusion of depth images and skeleton's key frames. Journal of Computer-Aided Design and Computer Graphics, 30(7): 1313-1320 (许艳, 侯振杰, 梁久祯, 陈宸, 贾靓, 宋毅. 2018. 权重融合深度图像与骨骼关键帧的行为识别. 计算机辅助设计与图形学学报, 30(7): 1313-1320) [DOI: 10.3724/SP.J.1089.2018.16771]

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI: 7444-7452

作者简介

尤凯军，男，硕士研究生，主要研究方向为动作识别和计算机视觉。E-mail：884098065@qq.com

侯振杰，通信作者，男，教授，主要研究方向为行为识别和机器视觉。E-mail：houzj@cczu.edu.cn

梁久祯，男，教授，主要研究方向为计算机视觉、图像处理和模式识别。E-mail：jzliang@cczu.edu.cn

钟卓锟，男，硕士研究生，主要研究方向为行为识别和计算机视觉。E-mail：327319110@qq.com

施海勇，男，硕士研究生，主要研究方向为行为识别和计算机视觉。E-mail：shihaiyong666@gmail.com