基于人体骨架特征学习的动作识别

林里浪，宋思捷，刘家瑛
（北京大学王选计算机研究所，北京 100080）

摘要：动作识别是计算机视觉研究中的一个基本但具有挑战性的问题。在过去的几年中，许多基于 RGB 视频的识别技术已经得到了巨大的发展，并取得了显著的成果。但是，处理 RGB 视频可能非常耗时。其中，在动作识别领域，人体骨架数据具有轻量级的特点，同时对人体外观、环境背景等信息具有不变性，因此，这种数据模态受到了越来越多的关注。然而，基于人体骨架的动作识别面临两个问题：人体骨架数据的噪声问题和数据标注的依赖问题。噪声问题是指骨架数据中存在噪声影响数据的准确性，而数据标注依赖问题则是指在监督学习中，需要大量的标签数据进行训练。本文针对人体骨架数据在采集中的噪声问题，提出了一种基于噪声适应的动作识别模型，设计了回归模型和生成模型充分利用不同场景下的噪声数据特点。并且针对人体骨架数据过于依赖标签数据，利用自监督学习方法，提出了一个基于多任务自监督学习的动作识别方法。
关键词：人体动作识别；骨架数据分析；特征学习
中图分类号：O422　**文献标识码：A**

Research on Skeleton Feature Learning based Human Action Recognition

Lilang Lin, Sijie Song, Jiaying Liu

(Wangxuan Institute of Computer Technology, 100080, China)

Abstract: Action recognition is a fundamental yet challenging problem in computer vision. In the past few years, many works have been developed on recognition based on RGB videos and achieved many significant results. However, processing RGB videos can be very time consuming. Another data modality, human skeletons, which represent a person by the 3D coordinate positions of skeletal joints, draw much attention due to the lightweight representations, the robustness to variations of viewpoints, appearances, and surrounding distractions. However, action recognition of skeleton data faces two problems: noise of skeleton data and dependence of data annotation. The problem of noise refers to the noise in skeleton data that affects the accuracy of data, while the problem of data annotation dependence refers to that the training requires lots of labelled data. To address the issue of action analytics from noisy skeletons which commonly appear in the real world, this paper proposes a noise-adaptation model to get rid of explicit skeleton noise modelling and reliance on skeleton ground truths. Regression-based and generation-based adaptation models are developed respectively according to whether the pairs of noisy skeletons are available. Besides, aiming at dependence of data annotation, with the self-supervised learning method on human skeleton data, this paper proposes an action recognition method based on multitask self-supervised learning.
Key words： human action recognition; skeleton-based action analysis; representation learning

1 引言

在信息时代，随着互联网、监控视频的发展，生产生活中存在海量视频图像数据，在这其中，人是主要视觉目标，对人体动作的分析与理解是视频图像分析理解应用的重点。其中，动作识别(Action Recognition)是计算机视觉领域一个基础且富有挑战性的问题。动作识别的需求迅速增长，促进了视频监控，人机交互和视频理解等领域的发展。

在过去的工作中，基于RGB视频数据的动作识别已经获得了较大发展且取得了许多显著的成果。然而，处理RGB视频数据非常耗时。另一种数据模态，人体骨架数据(Skeleton)，利用人体关节点的三维坐标来表示一个人体，实现了一种更加轻量级的表示方法。并且骨架数据对于视角的变换、人物的外貌以及周围的干扰具有较强的鲁棒性。此外，骨架序列可以视为人体动作的高级表征，吸引了许多研究人员研究基于人体骨架的动作识别[1][2][3]。

然而，目前基于人体骨架数据的研究受到了数据的限制。其中主要是数据的准确性以及数据的标注依赖。数据的准确性是在实验室环境和真实环境下，骨架数据由于采集环境不同导致数据质量有差异。实验室环境下的人体骨架数据质量较高，然而，实际应用中采集到的人体骨架多存在噪声问题。现有的基于人体骨架动作分析的工作多在实验室环境下的数据进行开发，未必能应对真实环境下噪声骨架的分析任务。而数据的标注依赖则是在现有的监督学习框架下，网络的训练需要大量的标签数据。这种情况通常是完全监督的方式进行训练，因此需要大量的标签数据进行训练。而标注训练数据是繁琐且昂贵的，针对这个问题，自监督学习(Self-Supervised Learning)被广泛应用，其利用数据自身的信息进行学习，从而摆脱了对于标签数据的依赖问题。然而，现有的自监督学习方法往往是基于单一的自监督任务进行学习，提取的特征较为单一，容易对于任务过拟合。

针对人体骨架动作识别中的这些问题，本文对人体骨架特征学习展开研究。针对人体噪声骨架问题，详细介绍基于噪声适应的人体骨架动作识别方法；并针对骨架数据的自监督学习任务，提出了基于多任务自监督学习的动作识别模型。

作者简介：林里浪（1999-），男（汉族），四川成都人，北京大学博士研究生，linlilang@pku.edu.cn。

2 预备知识

2.1 人体骨架数据

如图 1 所示，人体骨架数据是人体关节在空间中的三维坐标表示。主要通过姿态估计，深度相机采集，传感器采集等方式获得。随着硬件设备的发展以及姿态估计算法的优化，人体骨架数据开始在日常生活中发挥越来越大的作用，被广泛应用于各个方面。例如，在视频监控中，可以实时发现危险动作和违法行为；在健康监护中，可以监控老人小孩行为，预警跌倒摔跤；在人机交互中，可以手势控制智能设备；在运动分析中，可以姿态纠正动作评分等任务。

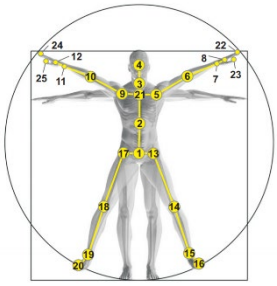

图 1 骨架数据示意图[4]

2.2 动作识别

人体骨架动作识别(Skeleton-Based Action Recognition)是基于骨架数据进行人体动作分类的技术。在获取到的骨架数据上，通过动作标签数据训练模型对其进行分类。动作的实施者可能为单人，如“理发”、“跳伞”，也可能为多人，如“皮划艇”、“握手”等。对于视频中的动作识别，一般而言，输入的视频通常包含一个动作，且该视频通常为切割好的视频(Trimmed video)，即视频不包含和该动作无关的视频帧。

动作识别是高层次视觉的典型问题，其准确性决定了后续任务如检测等问题的性能，是视频分析领域的基础。在动作识别领域，早期的经典方法基于局部特征描述进行识别，近来的工作多借助深度神经网络强大的表示能力，基于深层架构的学习来解决这一问题。

动作识别可被广泛应用于现实生活中，例如视频理解，运动康复以及行为预警等等。

2.3 人体骨架数据集

近年来随着人体骨架动作识别算法的研究发展，多个人体骨架数据集被相继提出并广泛使用，例如NW-UCLA[15]数据集，NTU[4]数据集和 PKU-MMD [17]数据集等数据集。其中 PKU-MMD [17]数据集是当前规模最大的多模态数据集，提供了包括人体骨架数据在内的多种数据模态（深度图、红外图等）下的动作视频。包含 1076 个未剪切的动作长序列，约 20000 个动作样本，累计包含 51 类动作。每个动作长序列为长度 3 至 4 分钟的视频，其中包含 20 个左右动作样本。该数据集分为两个部分：Part I 和 Part II，后者的骨架数据针对动作检测任务增强了动作发生的连续性，因此更加具有挑战性。

3 基于噪声适应的人体骨架动作识别

现有基于人体骨架的动作分析方法在均在标准人体骨架数据集上开发和评测，标准人体骨架数据通常在实验室环境下采集，视野清晰，设置规整，采集到的数据较为完整准确。实验室环境下的采集设置忽略了很多现实应用中存在的问题，如常见的遮挡、光线不足等。这些问题的存在会引起深度相机捕获人体骨架的质量降低，造成信息不准确、不完整等问题，形成人体骨架噪声。图 2 说明了实验室环境下和真实场景下人体骨架数据的差异，左边展示了实验室数据集的低噪声人体骨架，如SBU数据集[5]，NTU 数据集[4]，UTKinect 数据集[6]和 SYSU 数据集[7]，右边展示了真实场景下由于遮挡等问题产生的高噪声人体骨架。

针对上述问题，本文从噪声适应的角度解决基于噪声骨架的人体动作识别问题。将从以下三个方面对基于噪声适应的人体骨架动作识别展开介绍:

(1) 提出了新的骨架数据集 NSD (Noisy Skeleton Dataset)，包含了环绕式相机拍摄的同步人体骨架序列，模拟真实环境下由遮挡引起的骨架噪声。

(2) 提出了一种噪声等级估计方法，来定量衡量数据集中的噪声问题。

(3) 从噪声适应的角度为基于噪声骨架的人体动作识别提供解决方案，为充分利用现有骨架数据，分别提出了基于回归和基于生成的噪声适应模型。

3.1 NSD 数据集建立

本数据集一共收集了 1009 个长视频，每个视频时长为 1~2 分钟，包含大约 7 个动作片段，整个数据集包含 6952 个动作片段，涉及 41 类动作。在拍摄过程中，邀请了 13 名志愿者，每个志愿者参与到 4 个日常动作视频录制中，录制过程中志愿者的朝向不限。因此，在每次录制中，都存在由于严重遮挡引起的骨架噪声。

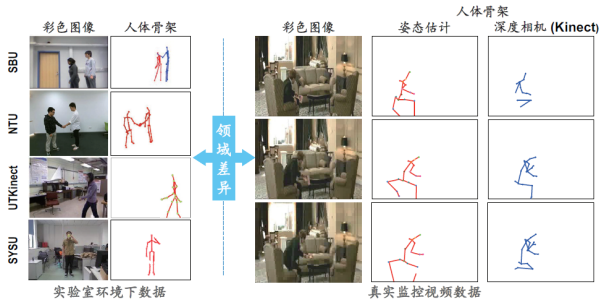

图 2 实验室环境下（左）和真实场境下（右）人体骨架数据存在差异

3.2 人体骨架噪声估计

本文设计了一种人体骨架噪声估计方法，以说明 NSD 数据集比现有数据噪声更加严重。由于很难获取噪声骨架对应的精确骨架信息，本文设计的噪声估计方法用相对的概念衡量人体骨架数据的噪声程度。随着姿态估计方法的发展，OpenPose 方法[8]能够从 RGB 图像中准确估计出 2D 人体骨架节点，可用作人体骨架数据的一种精确测量。对于 3D 人体骨架，可根据相机参数矩阵将其映射到 RGB 图像中，得到其在 RGB 图像上的二维坐标，通过对比上述两种数据，估计 3D 人体骨架的噪声程度。

表 1 NSD 数据集和 NTU[4]数据集噪声程度对比

	均值	方差
NTU-dx	0.258	0.064
NSD-dx	0.381	0.216
NTU-dy	0.181	0.058
NSD-dy	0.275	0.157

表 1 展示了 NSD 数据集的 3D 骨架投影出的 2D 骨架与姿态估计出的 2D 骨架数据在 x 轴和 y 轴上差值的均值与方差。同时与 NTU[4]数据集的结果做对比。可知 NSD 数据集的骨架数据噪声更加严重。

3.3 基于噪声适应的鲁棒特征学习

为了减少骨架噪声造成的动作识别性能损失，本文从噪声适应的角度出发，充分利用可用的人体骨架数据（如成对的噪声数据，以及非成对的不同噪声级别的骨架数据），以消除骨架噪声对动作识别任务的性能影响。分别尝试利用回归模型，将多

个不可靠的测量值收敛到未知的真实目标，以及利用基于对抗学习的生成模型，通过将数据适应到不同分布的方式来处理噪声数据或标签。

基于回归的噪声适应模型旨在从一个动作序列的多次观察中学习噪声鲁棒特征空间。尽管捕获真实精准的噪声数据难度大、成本高，但较易获得对同一动作序列的多次不可靠观察。

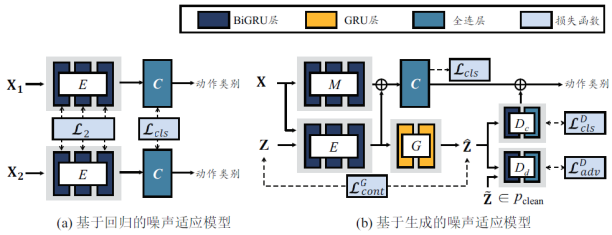

图 3 基于噪声适应的人体骨架动作识别模型

图 3(a)展示了基于回归的噪声适应模型的一个实例化，其中，X_1 和 X_2 分别为对同一动作序列的两次观察。采用双向 GRU (Bidirectional, GRU, BiGRU)网络建立编码器 $E(\cdot)$，将编码器第 l 层网络的隐层状态输出记为 $h_1^{l,t}$，$h_2^{l,t}$，则视频级别的动作序列可分别表示为 v_1^l，v_2^l，其为隐含层特征时域上的均值聚合。然后通过 l_2 范数约束视频级别的动作序列特征距离。由一层全连层构成的分类器 $C(\cdot)$ 将和视频级别的特征联合优化。此外，对 l_2 范数约束的距离项，采用参数 g 平衡其在优化中的作用，即总优化目标为：

$$L_{total} = gL_2 + L_{cls} \tag{1}$$

基于回归的噪声适应模型需要成对的噪声骨架数据。更进一步，对非成对噪声骨架的处理，采用基于生成模型的方式进行噪声适应。对于噪声程度比较低的骨架数据，对生成模型以对抗学习的方式对低噪声空间进行数据分布的建模，并将高噪声空间的特征适应到低噪声空间。

图 3(b)展示了基于生成的噪声适应模型的一个实例化，其中，X 和 Z 分别为高噪声和低噪声动作骨架，为简便，图中省略了一些损失函数。该模型的结构包括以下几个部分：主网络 $M(\cdot)$，分类器 $C(\cdot)$，编码器 $E(\cdot)$ 和解码器 $G(\cdot)$，判别器 $D_d(\cdot)$ 和 $D_c(\cdot)$。基于对抗学习的思想，解码器 $G(\cdot)$ 生成骨架以迷惑判别器，从而约束编码器 $E(\cdot)$ 学习噪声抑制的动作序列特征。判别器的组成包含两部分：$D_d(\cdot)$ 为判断输入是否来自真实分布的二值分类器，另一部分 $D_c(\cdot)$ 为 C 路分类器，对输入的类别概率分布建模。为了进一步提升测试阶段的动作识别准确度，将模型中两个分类器 $C(\cdot)$ 和 $D_c(\cdot)$ 的结果整合，最终的动作识别结果为：$C(M(X)+E(X))+D_c(G(E(X)))$。

3.4 实验结果

表 2 展示了不同方法在 NSD 数据集上的动作识别准确率。对于基于回归的噪声适应模型，通过将多个同时观察到的人体骨架动作序列映射到噪声鲁棒的噪声空间，本文提出的基于回归的噪声适应模型(R-NAN)能够有效将基础网络(Baseline)的性能在不同设置下分别提升 4.6%和 5.9%。对于基于生成的噪声适应模型，通过对低噪声人体骨架数据的数据分布建模，本文提出的基于生成的噪声适应模型(G-NAN)能够有效将基础网络(Baseline)的性能分别提升 4.8%和 1.7%。

实验中同时测试了现有方法在 NSD 数据集上的性能，以分析现有方法是否对噪声骨架的动作识别鲁棒，表 2 展示了其结果。对于基于注意力机制的模型[9][10]，人体骨架中的噪声扰动易误导注意力权重的分配，因此，基于注意力机制的模型在噪声人体骨架的动作识别上有明显的性能退化。方法 TPN[11]和 VA-LSTM[12]由于噪声干扰，也未能实现有效的人体骨架特征提取。尽管方法 Denoised-LSTM[13]显式地使用姿态编码的方式将人体骨架去噪，但这一过程过度平滑了人体骨架，破坏了利于动作识别的细节。

和上述方法相比，本文提出的基于回归的噪声适应方法和基于生成的噪声适应方法均在噪声骨架的动作识别上取得了显著提升。

表 2 不同方法在 NSD 数据集上的动作识别准确率

方法	Cross-Subject	Cross-View
STA-LSTM[9]	44.3	28.6
TPN[11]	46.9	29.7
VA-LSTM[12]	50.0	34.5
Denoised-LSTM[13]	38.1	26.1
Baseline	50.7	34.6
R-NAN	55.3	40.5
G-NAN	55.5	36.3

4 基于多任务自监督学习的动作识别

上一节解决了骨架数据中的噪声问题，然而在训练中依然需要大量的标签数据。为了减轻训练对于标注数据的依赖，本文提出了基于多任务自监督学习的方法，进行无监督的特征学习。

自监督任务旨在通过大量的无标签数据提取数据特征，通过相应的自监督任务生成标签数据，通过监督的方法提取出有意义的特征，并辅助目标任务的学习。然而，之前的自监督学习工作从单个任务中学习可能会导致对特定任务的过度拟合。因此，从单一的重建工作中学习到的特征对于识别骨架序列来说可能不够具有区分性和泛化性。

针对上述问题，本文介绍一种通过同时优化多个任务的自监督学习方法。通过探索组合不同的任务，使特征表示更加多样化并描述信息的不同方面。在本文中，设计了三个任务，运动预测的生成任务，解决拼图的分类任务，以及基于骨架数据变换的对比学习。这些自监督任务旨在从运动预测中学习骨骼动力学特征，通过解决拼图来提取时间信息，并通过对比学习进一步规范特征空间。

本文还在实验中提供了全面的评估和分析，以证明提出的自监督学习方法的优越性。

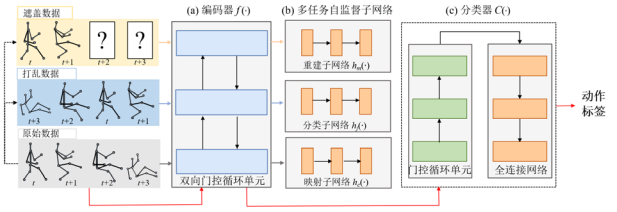

图 4 基于多任务自监督学习的人体骨架动作识别模型

4.1 多任务自监督学习

为了学习到更加泛化且鲁棒的特征，本文考虑了多种自监督任务，例如动作预测，拼图任务以及对比学习。本文希望通过动作预测学习到骨架数据的动态信息，通过拼图任务学习到时域特征。最后，利用对比学习来规范特征空间。

图 4 展示了本文所采用的模型框架。这些任务共享编码器 $f(\cdot)$ 并且采用不同的子网络结构。其中，重建子网络 $h_m(\cdot)$ 利用编码器 $f(\cdot)$ 提取的特征重建原始数据。分类子网络 $h_j(\cdot)$ 输入打乱数据的特征预测打乱的顺序。映射子网络 $h_c(\cdot)$ 输入变换数据和原始数据，拉进正例特征而推远负例特征。最后，分类器 $C(\cdot)$ 输入预训练好的编码器 $f(\cdot)$ 提取的特征进行动作分类训练。

4.2 动作预测

动作预测是给定过去的骨架序列，通过对人体骨架动态建模预测未来的动作。本文应用了基于循环神经网络的自编码器来完成动作预测。输入原始骨架数据的部分提取特征，再 重建出原始骨架数据。其输入提取的特征，然后重建出原始骨架数据。

假设输入的原始骨架数据为 $X^i = \{x_1^i, \cdots, x_T^i\}$，被遮盖后的数据为 $X_m^i = \{x_1^i, \cdots, x_{T\phi}^i | T\phi \quad T\}$。因此预测的骨架数据为 $\hat{X}_m^i = h_m(f(X_m^i))$，这里 $\hat{X}_m^i = \{\hat{x}_{T\phi+1}^i, \cdots, \hat{x}_T^i\}$。本课题采用均方损失函数(MSE)来训练网络。

$$L_m = \sum_{i=1}^{N} \sum_{t=T\phi+1}^{T} \left\| \hat{x}_t^i - x_t^i \right\|_2^2 \tag{2}$$

4.3 拼图任务

如图 5 所示，拼图任务旨在从打乱的序列中预测出正确的排列方式。为了从原始骨架数据中生成打乱的数据，将每个序列分为 3 个子片段。将 3 个子片段随机打乱，有 6 种打乱方法。然后将打乱的数据输入网络中预测打乱的片段的正确顺序。

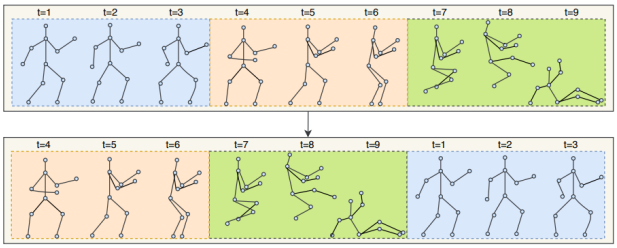

图 5 骨架数据分段打乱方法示意图

通过共享的编码器 $f(\cdot)$，采用一个分类子网络，$h_j(\cdot)$ 来获得了拼图任务的分类结果。特别地，本文采用多层感知机作为分类头，利用交叉熵损失作为分类损失。令 X_j^i 为被打乱后的数据，y^i 为打乱的标签。

$$L_j = -\sum_{i=1}^{N} y^i \log h_i(f(X_j^i)) \quad (3)$$

4.4 对比学习

为了进一步规范特征学习并鼓励网络学习固有表示，本文算法通过将变换后的数据映射到特征空间来采用对比学习。受 SimCLR[16]的启发，网络通过最大化相同原始数据的转换模态之间的余弦相似度来学习表示。对于每个原始样本，本文算法将考虑多次变换。具体来说，随机抽取 N 个样本并应用 M 种变换操作来获得 NM 个变换样本。然后对于每个原始样本，可以与其转换后的样本构建 M 对正对，并与其他样本构建负对。

投影子网络 $h_c(\cdot)$ 旨在将编码序列映射到特征空间。设 $z_1, z_2, \cdots, z_N$ 是 从编码器 $f(\cdot)$ 和映射子网络 $h_c(\cdot)$ 的输出中提取的特征 的原始数据。并令 $z_1^t, z_2^t, \cdots, z_N^t$ 是来自原始序列的变换样本的特征，然后利用均值整合转换后的特征。

本文使用 $sim(\mathrm{x},\mathrm{y}) = \dfrac{x^T y}{\|x\|_2 \|y\|_2}$ 作为余弦相似度，而损失函数采用：

$$L_c = -\sum_{t=1}^{M}\sum_{k=1}^{N} \log \frac{\exp(sim(z_k, z_k^t))}{\sum_{j=1}^{N} \exp(sim(z_k, z_j))} \quad (4)$$

在实际应用中，本文采用两种变换方式，时域的遮盖与时域的打乱，这两种变换后的数据分别作为动作预测和拼图任务的输入。

4.5 实验结果

在无监督学习设定中，编码器单独使用自监督任务训练，然后固定住编码器的参数，利用其提取的特征训练一个线性分类器。利用该线性分类器度量特征的质量。

本文在 NW-UCLA[15]数据集，NTU[4]数据集和 PKU-MMD [17]数据集上进行测试。NW-UCLA[15]数据集含有 1494 个视频数据，包含 10 个类别，每个骨架数据 20 个关节点。NTU[4]数据集含有 56578 个视频数据，由 60 类动作组成，每个数据有 25 个关节。PKU-MMD [17]数据集有 20000 个动作样本，累计包含 51 类动作，每个骨架数据包含 25 个节点。

在表 3 和表 4 中，展示了自监督学习方法的结果。其中，基础网络(RandU)为随机初始化网络作为编码器的基线方法。本文的方法(MS^{2}L)在大多数设置中比基础网络(RandU)和 LongT GAN[14]获得更好的性能。LongT GAN[14]只使用了重建自监督任务作为预训练，所以对特征的提取能力较为单一。而本文通过多任务的组合提取到了更加丰富的特征。

表 3 不同方法在 NW-UCLA[15]和 NTU[4]数据集上的动作识别准确率

方法	NW-UCLA[15]	NTU[4]
RandU	60.61	41.10
LongT GAN[14]	74.30	52.14
MS^{2}L	76.81	52.55

表 4 不同方法在 PKU-MMD[17]数据集上的动作识别准确率

方法	Part I	Part II
RandU	51.6	28.4
LongT GAN[14]	67.7	25.9
MS^{2}L	64.8	27.6

5 结论

在基于人体骨架的动作分析方法中，面临两个问题。分别是骨架数据的噪声问题和监督训练的标签依赖问题。人体骨架噪声是人体骨架采集中难以避免的问题，而现有针对人体骨架的动作识别方法均未考虑噪声对人体动作识别的影响。针对这一问题，本文提出了一种基于噪声适应的动作识别模型，以减小噪声在训练过程中和测试过程中对模型的影响，同时避免噪声建模和对真实非噪声数据的依赖。本文通过实验验证了基于噪声适应的动作识别方法的有效性。同时，为了减少数据对于标签数据的依赖，针对自监督骨架特征学习问题，本文提出了一种基于多任务自监督学习的骨架数据动作识别方法，包括动作预测，拼图任务以及对比学习。本文利用动作预测对骨架数据的低级信息进行建模，通过拼图任务提取到数据的时域特征，同时，利用对比学习对特征空间进行规范使得特征更具可分性且更加泛化。通过实验，证明了本文的方法的优越性。通过这两个模型，本文解决了骨架数据训练中面临的两个问题，并获得了较好的性能。

参考文献

[1] Xikun Zhang, Chang Xu, Dacheng Tao. Context Aware Graph Convolution for Skeleton-Based Action Recognition[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2020, 14333–14342.

[2] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong

Wang, Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2020, 143–152.

[3] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, Nanning Zheng. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2020, 1112–1121.

[4] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang. NTU RGB+ D: A Large Scale Dataset for 3D Human Activity Analysis[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2016, 1010–1019.

[5] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop. 2012, 28– 35.

[6] Lu Xia, Chia-Chih Chen, Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop. 2012, 20–27.

[7] Jian-Fang Hu,Wei-Shi Zheng, Jianhuang Lai, Jianguo Zhang. Jointly learning heterogeneous features for RGB-D activity recognition[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2015, 5344–5352.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2017, 7291–7299.

[9] Sijie Yan, Yuanjun Xiong, Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Proc. AAAI Conference on Artificial Intelligence. 2018, 7444–7452.

[10] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jiaying Liu. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data[C]. Proc. AAAI Conference on Artificial Intelligence. 2017, 4263–4270.

[11] Yueyu Hu, Chunhui Liu, Yanghao Li, Sijie Song, Jiaying Liu. Temporal Perceptive Network for Skeleton-Based Action Recognition[C]. Proc. British Machine Vision Conference. 2017, 1–12.

[12] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, Nanning Zheng. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data[C]. Proc. IEEE Int'l Conference on Computer Vision. 2017, 2117–2126.

[13] Girum G Demisse, Konstantinos Papadopoulos, Djamila Aouada, Bjorn Ottersten. Pose encoding for robust skeleton-based action recognition[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop. 2018, 188–194.

[14] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, Zhefeng Gong. Unsupervised Representation Learning With Long-Term Dynamics for Skeleton Based Action Recognition[C]. Proc. AAAI Conference on Artificial Intelligence. 2018, 2644–2651.

[15] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, SongChun Zhu. Cross-view action modeling, learning and recognition[C]. Proc. IEEE Conference on Computer Vision and Pattern Recognition. 2014, 2649–2656.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton. A simple framework for contrastive learning of visual representations[C]. Proc. Int'l Conference for Machine Learning. 2020, 1597–1607.

[17] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, Jiaying Liu. PKU-MMD: A large scale benchmark for skeleton-based human action understanding[C]. Proc. the Workshop on Visual Analysis in Smart and Connected Communities. 2017, 1–8.