第 58 卷 第 24 期/2021 年 12 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于 2D 转 3D 骨架的多特征融合实时动作识别

任国印[1,2]，吕晓琪[1,2,3*]，李宇豪[2]

[1]内蒙古科技大学机械工程学院，内蒙古 包头 014010；

[2]内蒙古科技大学信息工程学院，内蒙古 包头 014010；

[3]内蒙古工业大学，内蒙古 呼和浩特 010051

摘要 提出了一种基于二维(2D)转三维(3D)骨架的实时检测双分支子网络，可实现 2D 骨架关键点的 3D 估计和 2D、3D 骨架特征融合的人体 3D 动作识别。在检测过程采用 OpenPose 框架实时获取视频中人体骨架的 2D 关键点坐标。在 2D 转 3D 骨架估计过程中，设计了一种输入为难样本且具有反馈功能的孪生网络。在 3D 动作识别过程中设计了一种 2D、3D 骨架特征双分支孪生网络，以完成 3D 姿态识别任务。在 Human3.6M 数据集上训练 3D 骨架估计网络，在基于欧拉变换的 NTU RGB+D 60 多视角增强数据集上训练骨架动作识别网络，最终得到的 3D 骨架动作识别交叉受试者准确率为 88.2%，交叉视野准确率为 95.6%。实验结果表明，该方法对 3D 骨架的预测精度较高，且具有实时反馈能力，可适用于实时监控中的动作识别。

关键词 图像处理；三维骨架估计；人体动作识别；多分支网络；多特征融合

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2410010

Multi-Feature Fusion Real-Time Action Recognition Based on 2D to 3D Skeleton

Ren Guoyin[1,2], Lü Xiaoqi[1,2,3*], Li Yuhao[2]

[1] *School of Mechanical Engineering, Inner Mongolia University of Science & Technology, Baotou, Inner Mongolia* 014010, *China*;

[2] *School of Information Engineering, Inner Mongolia University of Science & Technology, Baotou, Inner Mongolia* 014010, *China*;

[3] *Inner Mongolia University of Technology, Huhhot, Inner Mongolia* 010051, *China*

Abstract We propose a real-time detection binary sub network based on two-dimensional (2D) to three-dimensional (3D) skeleton, which can realize 3D estimation of key points of 2D skeleton and human 3D motion recognition based on 2D and 3D skeleton feature fusion. In the detection process, OpenPose framework is used to obtain the 2D key point coordinates of human skeleton in video in real time. In the process of 2D to 3D skeleton estimation, a siamese network with difficult input samples and feedback function is designed. In the process of 3D motion recognition, a two branch siamese network of 2D and 3D skeleton features is designed to complete the task of 3D pose recognition. The 3D skeleton estimation network is trained on the Human3.6M data set, and the skeleton action recognition network is trained on the NTU RGB+D 60 multi view enhancement data set based on Euler transform. Finally, the accuracy of cross subjects and accuracy of cross views are 88.2% and 95.6%. Experimental results show that the method has high prediction accuracy for 3D skeleton and real-time feedback ability, and can be applied to action recognition in real-time monitoring.

Key words image processing; three-dimensional skeleton estimation; human action recognition; multi branch network; multi-feature fusion

收稿日期：2021-01-18；**修回日期**：2021-02-09；**录用日期**：2021-03-09

基金项目：国家自然科学基金(61771266，81571753)、包头市青年创新人才项目(0701011904)

通信作者：*1712152231@qq.com

OCIS codes 100.3010; 100.4994; 100.4996; 110.6880

1 引 言

目前，基于摄像机的实时人体动作识别技术突破了动作识别研究领域的诸多难点。实际生活中，密集人群监控、人机交互和自动驾驶等领域也尝试使用人体行为识别技术监控目标人员的活动，且都起到了至关重要的作用。虽然对动作识别的研究已经取得了长足的进展，但在实时视频的动作识别中仍存在一些问题，如动作被遮挡时无法识别、识别速度慢等问题，而三维(3D)动作识别在避免遮挡、减少背景干扰、动作捕捉准确度等方面具有明显优势，因此，人们逐渐将动作识别的研究重点由二维(2D)空间转向 3D 空间。

将 2D 空间转换到 3D 空间的方法主要有传统的图像处理方法和深度学习方法两类。传统的空间变换方法，如通过轮廓[1]及形状上下文[2]等特征筛选描述符、通过边缘方向直方图[3]获取行为特征、加权尺度不变特征变换[4]、局部特征匹配[5]、光流场的点云变换[6]或深度 RGB(RGB-D)图像变换方法[7]依赖于人工设计的抽取器，且需要复杂的参数调整过程，存在一定的局限性。Chen 等[8]提出了一种从单张 RGB(Red，Green，Blue)图像中估计 3D 人体姿态的方法，将深度特征与 2D 关键点数据结合，完成基于信息融合的网络学习机制，进而预测出 3D 姿态。该方法不仅适用于室内动作识别，还可以应用在户外行人动作的预测中。Martinez 等[9]提出了一种基于深度学习前馈网络的 3D 关节预测方法，该方法通过 $N\times N$ 的距离矩阵表示 2D 和 3D 人体姿态，并用距离矩阵回归得到更精确的姿态。Moreno-Noguer[10]提出了一种从单张图像中估计 3D 人体姿态的方法，进一步提高了 3D 姿态的估计精度。RGB-D 相机在户外场景下受阳光的干扰经常出现故障，且其体积大、图像噪声高、分辨率低、摄程有限，因此，人们尝试用其他图像采集设备代替 RGB-D 相机。Mehta 等[11]提出了一种基于 Kinect 全卷积的 3D 人体实时姿态估计方法，可以稳定产生骨架全局姿态并分析关节角度，取得了较高的实时估计精度。但 Kinect 采集的关节深度图像通常存在噪声，容易引起模型过拟合。基于回归的方法没有很好地利用姿态中的深度信息，因此，Sun 等[12]提出了一种基于实时检测的方法，用骨架代替关节，并借助 OpenPose[13]的实时检测能力，大大提高了 Human3.6M 数据集的 3D 泛化能力。Cippitelli 等[14]提出了一种利用 Kinect 的人体动作识别网络，该方法用提取的关键姿态组成特征向量，并用多类支持向量机进行分类，具有很好的鲁棒性，但处理图像前需要进行降噪处理。之后，Aubry 等[15]提出了一种基于实时检测的动作识别方法。OpenPose 能使骨架的获取更加便捷，也解决了利用 Kinect 获取深度图像时出现的噪声问题。OpenPose 的骨架数据被编码成 RGB 三通道的图像，将骨架运动序列转换成 RGB 图像后作为神经网络的输入，通过训练图像分类神经网络识别动作。该方法忽略了 3D 骨架的深度信息，在 2D 骨架被遮挡和覆盖时会影响最终的分类结果。Pham 等[16]提出了一种基于深度学习的多任务 3D 人体姿态估计框架，并用简单的摄像机从 RGB 传感器中识别人体 3D 动作。此外，从人体骨架 3D 关节位置提取的特征具有显著性，可以有效地用于动作识别，且该系统在低延迟情况下的检测精度优于现有的同类方法。Yang 等[17]提出了一种基于双特征双运动网络的骨架动作识别，发现 2D 姿态和 3D 姿态估计任务本质上具有相互纠缠、信息互补、特征共享的特性，因此，特征信息融合在 3D 姿态识别中发挥着重要作用。

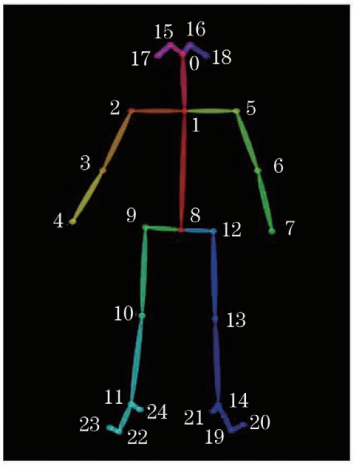
本文借助具有反馈功能的难样本孪生网络实现 2D 骨架关键点的 3D 估计，同时借助双分支孪生网络实现 2D、3D 骨架特征融合的人体 3D 动作识别。其中，2D 骨架关键点的 3D 估计可以提升骨架关键点的空间特征精度，而 2D、3D 骨架特征的融合可弥补低维空间因遮挡导致的动作特征混淆问题。结合这两种方法可提升 3D 骨架动作的预测精度，降低骨架动作识别的误判率。此外，在视频监控处理方面孪生网络的反馈能力也可实现实时的 3D 动作识别。

2 网络设计

2.1 骨架姿态探测器

图像处理领域一般用运动捕捉系统获取 3D 深度图像，常见的运动捕捉系统包括 RGB-D 相机和 Kinect 深度传感器，但这两种系统只能在有限的视野和距离内操作。此外，低成本深度传感器在强光下工作，特别是在阳光下工作时会伴随着大量图像噪声。因此，采用一个实时的多人 2D 姿态探测器 OpenPose 生成 25 个 2D 人体骨架关键点，具体的

关键点编号和部位如图 1 所示。相比 RGB 图像，3D 骨架对光照条件的变化不敏感，数据更精确，可减少噪声干扰。此外，从人体骨架 3D 关节位置提取的特征具有显著性，可以有效地用于动作检测与识别。



0	nose	13	L-knee
1	neck	14	L-ankle
2	R-shoulder	15	R-eye
3	R-elbow	16	L-eye
4	R-wrist	17	R-ear
5	L-shoulder	18	L-ear
6	L-elbow	19	L-bigtoe
7	L-wrist	20	L-smalltoe
8	midhip	21	L-heel
9	R-hip	22	R-bigtoe
10	R-knee	23	R-smalltoe
11	R-ankle	24	R-heel
12	L-hip		

图 1　OpenPose 采集的 2D 人体骨架关键点部位及编号

Fig. 1　Key points and numbers of the 2D human skeleton collected by OpenPose

2.2　3D 骨架姿态估计

目前，基于骨架的动作识别已经得到了广泛应用。在基于视频的动作识别中，很多算法用 2D 人体骨架关键点作为研究对象。OpenPose 可实时采集 2D 人体骨架坐标，如果这些 2D 骨架附带动作标注信息，则可通过设计一个高性能的深度学习网络实时识别这些 2D 人体动作。但 2D 关键点易出现重叠、视觉错位和遮挡等问题，忽略了人体动作的时空信息，在提高系统的识别精度方面还存在不足。因此，人们逐渐将研究重点转移到 2D 骨架深度信息的获取中。要实现 3D 空间中人体骨架关键点的估计，需要先得到人体关键点的 3D 坐标，从而更精确地预测和识别人体动作。因此，提出了一种基于 3D 骨架姿态探测器的骨架估计方法，以达到 RGB-D 图像采集传感器的效果。实验的研究目的是从给定实时监控或离线 RGB 视频中截取一段视频，然后将这段视频中估计的 3D 人体骨架动作按照时间顺序播放出来。其中，每个 3D 人体骨架动作均由动作骨骼的 25 个关键点构成。时间段 $t=(t_0,\cdots,t_n)$ 范围内估计的一系列 3D 姿态 $\boldsymbol{X}^{\mathrm{P}}=(\boldsymbol{p}_0,\cdots,\boldsymbol{p}_n)$，其中，$\boldsymbol{X}^{\mathrm{P}}\in\boldsymbol{p}^{3\times N}$，$N$ 为动作的关键点数量。由于选用 OpenPose 的 BODY-25 骨架关键点模型，因此 $\boldsymbol{X}^{\mathrm{P}}\in\boldsymbol{p}^{3\times 25}$。

通过设计一个双分支孪生监督学习模型将生成的 $\boldsymbol{X}^{\mathrm{P}}$ 作为神经网络的输入，以预测动作类型。对于视频中的每一帧 RGB 图像，预测其在 3D 空间中对应的 3D 人体骨架关键点的立体分布。人体骨架在 3D 空间的像素分布可表示为 $\boldsymbol{P}_{3\mathrm{D}}\in\boldsymbol{p}_i^{3\times N}$，首先，借助 OpenPose 人体骨架 2D 姿态检测器获取骨架 2D 关键点位置像素矩阵 $\boldsymbol{P}_{2\mathrm{D}}\in\boldsymbol{p}_i^{2\times N}$。通过设计一个双分支孪生监督学习模型建立一个转换函数 $X_{\mathrm{change}}(\boldsymbol{P}_{2\mathrm{D}})$，以恢复骨架的 3D 位置，可表示为

$$\boldsymbol{P}_{3\mathrm{D}}=X_{\mathrm{change}}[\boldsymbol{P}_{2\mathrm{D}},\ \boldsymbol{p}(\boldsymbol{c})],\tag{1}$$

式中，$\boldsymbol{p}(\boldsymbol{c})=\{\boldsymbol{p}_1,\cdots,\boldsymbol{p}_n\}$ 为双分支监督学习模型的可训练参数，$X_{\mathrm{change}}()$ 为一个损失函数，通过训练优化该损失函数，使 3D 姿态 $\boldsymbol{P}_{3\mathrm{D}}$ 和真实骨架 3D 数据的损失最小，可表示为

$$d_{\min}=\arg\sum_{n=1}^{N}X_{\mathrm{Loss}}[X_{\mathrm{change}}(\boldsymbol{P}_{2\mathrm{D}}),\ \boldsymbol{P}_{3\mathrm{DT}}],\tag{2}$$

式中，$d_{\min}$ 为预测 3D 骨架和真实 3D 骨架的最小距离，$\boldsymbol{P}_{3\mathrm{DT}}$ 为本地真实 3D 骨架姿态位置坐标，X_{Loss} 为预测 3D 骨架坐标的损失函数。

深层网络在提取高维特征方面的表现优越，但将其用于 2D 人体骨架的坐标数据时，会出现过拟合问题。为了避免这类问题，需要设计合适的卷积层数减轻因卷积层数过深导致的梯度消失和梯度爆炸现象。因此，提出了一种改进的双分支孪生网络，其结构如图 2 所示。该网络的每个分支都包括卷积层(Convolutional layer)、线性层(Linear layer)、批量标准化(BN)、Dropout 层、激活函数(SELU)和全连接层(FC)。第一个网络分支用 Human3.6M 数据集中自带的真实 3D 坐标作为输入；第二个网络分支用 OpenPose 采集的 2D 人体骨骼可视化关键点作为输入。最后对两个分支的输出进行加权处理，在两个分支中导入真实 2D 坐标，以完成监督学习，从而提高输出深度信息的精度，提升训练模型的

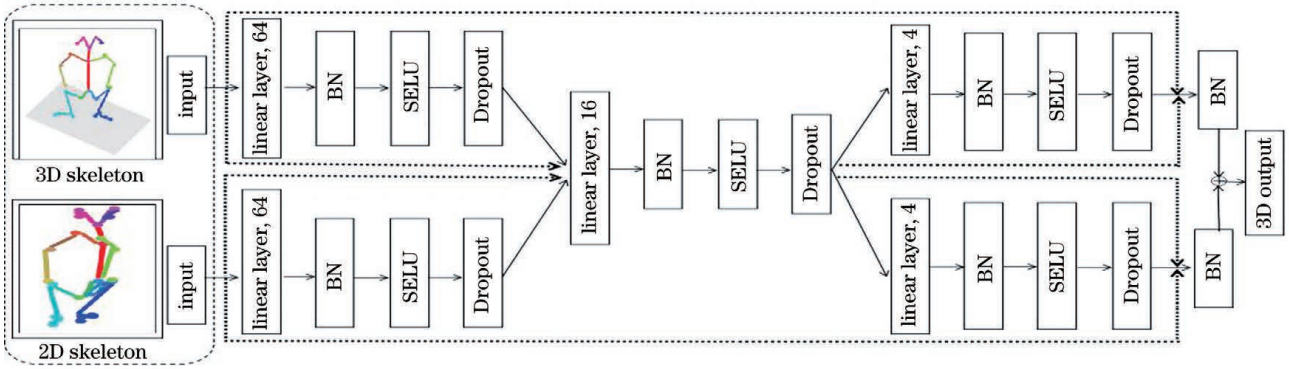


图 2　训练 3D 骨架估计器的网络结构

Fig. 2　Network structure for training 3D skeleton estimator

泛化能力，使模型的鲁棒性更强。

2.3　Human3.6M 数据集的数据加强

基于深度学习方法的效果虽然比人工提取特征的方法更好，但缺乏大量的数据进行训练。Human3.6M 数据集[18]包含 360 万个精确的 3D 人体姿态，数据集中典型的人类活动包括拍照、打电话、摆姿势、打招呼、吃饭等，这些姿态均由 5 名女性和 6 名男性受试者在 4 种不同视角下采集得到。该数据集附加了同步于图像的人体运动深度数据，可作为评估人体姿态 3D 估计模型的标准数据集，但该数据集的有效数据覆盖率还有待提高。因此，通过变换观察角度和人体姿态的样本量进一步提升和加强大型数据集的训练潜力，也是本研究的创新点之一。

已有的骨架数据集通常是在一定条件下生成的，对于Human3.6M数据集来说，其骨架视频数据是在 4 个视角下采集的。由于骨架视频的采集角度有限，用其训练神经网络时只能在特定的 4 个视觉角度下优化网络参数，导致模型对观察视角的泛化能力较差。而深度卷积模型的优越表现往往基于足够多的泛化样本进行训练优化，以避免过拟合问题。因此，需要通过数据增强方式扩充样本基数，进而达到扩大样本数量的目的。

通过数据增强技术扩充骨架数据，通过增加采集样本的观察视角提高样本的多样性，增强模型的泛化能力。具体实现方法：用欧拉角公式在 3D 笛卡儿空间完成变换[19]，使空间坐标在笛卡儿空间内沿指定方向进行旋转。3D 物体在空间内完成的任意一次旋转，都可理解为笛卡儿空间三个坐标方向的轴旋。若 3D 物体在 x、y、z 坐标轴上进行逆时针转动，且其旋转角度分别为 θ_1、θ_2、θ_3，则旋转变化矩阵可表示为

$$\boldsymbol{R}_x(\theta_1)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\theta_1 & -\sin\theta_1\\ 0 & \sin\theta_1 & \cos\theta_1\end{bmatrix},\ \boldsymbol{R}_y(\theta_2)=\begin{bmatrix}\cos\theta_2 & 0 & \sin\theta_2\\ 0 & 1 & 0\\ -\sin\theta_2 & 0 & \cos\theta_2\end{bmatrix},\ \boldsymbol{R}_z(\theta_3)=\begin{bmatrix}\cos\theta_3 & -\sin\theta_3 & 0\\ \sin\theta_3 & \cos\theta_3 & 0\\ 0 & 0 & 1\end{bmatrix}。\tag{3}$$

空间物体的每一次旋转操作都可认为是(3)式中 3 个基本矩阵按照一定顺序的乘积，旋转矩阵 $\boldsymbol{R}$ 可表示为

$$\boldsymbol{R}=\boldsymbol{R}_x(\theta_1)\boldsymbol{R}_y(\theta_2)\boldsymbol{R}_z(\theta_3)。\tag{4}$$

先绕 x 轴旋转角度 θ_1，再绕 y 轴旋转角度 θ_2，最后绕 z 轴旋转角度 θ_3 后，得到的旋转矩阵可表示为

$$\boldsymbol{R}_{xyz}(\theta_1,\theta_2,\theta_3)=\begin{bmatrix}\cos\theta_1\cos\theta_2 & \cos\theta_1\sin\theta_2-\sin\theta_1\cos\theta_3 & \cos\theta_1\sin\theta_2\cos\theta_3+\sin\theta_1\sin\theta_3\\ \sin\theta_1\cos\theta_2 & \sin\theta_1\sin\theta_2\sin\theta_3 & \sin\theta_1\sin\theta_2\cos\theta_3-\cos\theta_1\sin\theta_3\\ -\sin\theta_2 & \cos\theta_2\sin\theta_3 & \cos\theta_2\cos\theta_3\end{bmatrix}。\tag{5}$$

如果 3D 空间 x、y、z 三个坐标都加上欧拉旋转，数据集原有的采集角度将会得到大幅度扩充。使用 Human3.6M 标准数据集完成骨架的 3D 估计，如果将骨架的 3D 估计结果与 NTU RGB+D 60 对应的 3D 骨架估计数据同时进行欧拉样本增强并作为动作识别网络的输入样本，可提升该网络在动作识别上的精确性和泛化性[20-22]。欧拉样本增强的具体步骤：使数据在 x 轴和 y 轴固定，沿 z 轴旋转；

使数据在 x 轴和 z 轴固定，沿 y 轴旋转；使数据在 z 轴和 y 轴固定，沿 x 轴旋转。所用样本集均基于欧拉角旋转公式按照这三种方法完成样本视角多样化的扩充。

2.4　基于多特征融合网络的动作识别

本方法中的多特征融合多分支深度神经网络以2D骨架特征、3D骨架空间特征及骨架的真实动作信息作为输入，通过多特征融合的双分支卷积神经网络实现3D骨架行为识别。该网络通过叠加方式在3D姿态估计器上改变训练时的损失函数，其中，人体骨架3D估计的网络输出为多特征融合网络的一个输入。整合3D姿态空间转换器和动作识别器，实现RGB视频流中人体骨架的3D重建与实时动作识别。双分支神经网络可充分融合两对特征，两个分支的输入主要采用Trihard loss的难样本输入思想。传统双分支孪生网络的样本输入是随机抽取训练数据中的3张图像，且大多数样本是易于识别的简单样本对，不利于网络学习更好的表征。

大量研究结果表明，用更难识别的样本进行训练可提高网络的分类精度。样本输入分别为NTU RGB+D 60中自带的真实3D骨架坐标(①)、用NTU RGB+D 60数据集通过双分支网络估计的3D骨架坐标(②)、用OpenPose采集的2D骨架坐标(③)。三种特征在配对融合前看似相互独立，但内部存在很多关联。由于2D和3D信息都是来自同一个样本的估计与识别，因此将多特征充分融合可提升本网络的识别精度。目前大部分融合方法是在输出部分通过加权完成简易融合[23-26]，不能学习特征间潜在的互异性和相似性，因此，设计了一种可互换权重的双分支孪生网络，以充分挖掘融合特征。

为了融合人体骨架的2D和3D动作特征，设计了一种双分支孪生网络，该网络的结构如图3所示。网络包括三种输入特征：骨架的2D特征、估计的骨架3D特征和真实的骨架3D特征。骨架的2D特征由OpenPose采集的2D骨架坐标和对应的动作标签组成；3D特征由NTU RGB+D 60数据集通过动作估计生成的3D骨架坐标及其对应的动作标签组成；真实的骨架3D特征及其动作类别标签可用作3D特征的校正样本。在预处理阶段将骨架的3D特征和真实3D坐标通过欧拉角转换进行多视角扩充；将三种输入样本配成两对，一对是估计的3D坐标和真实3D坐标，另一对是估计的3D坐标和OpenPose采集的2D坐标，前提是保证每一次样本配对来自同一个人，以学习样本的潜在泛化信息。实验评估结果表明，估计的3D坐标和真实的3D坐标非常接近，可作为相似特征的正样本对输入；估计的3D坐标和OpenPose采集的2D坐标是区别较大的特征样本对输入，可认为是差别较大的正样本对，可加大网络样本输入难度，增加泛化性因子。因此，可将2D骨架的位置信息充分融合到3D骨架信息中，进而深层次挖掘骨架在空间维度上的潜在联系，提升骨架动作识别的精确度。

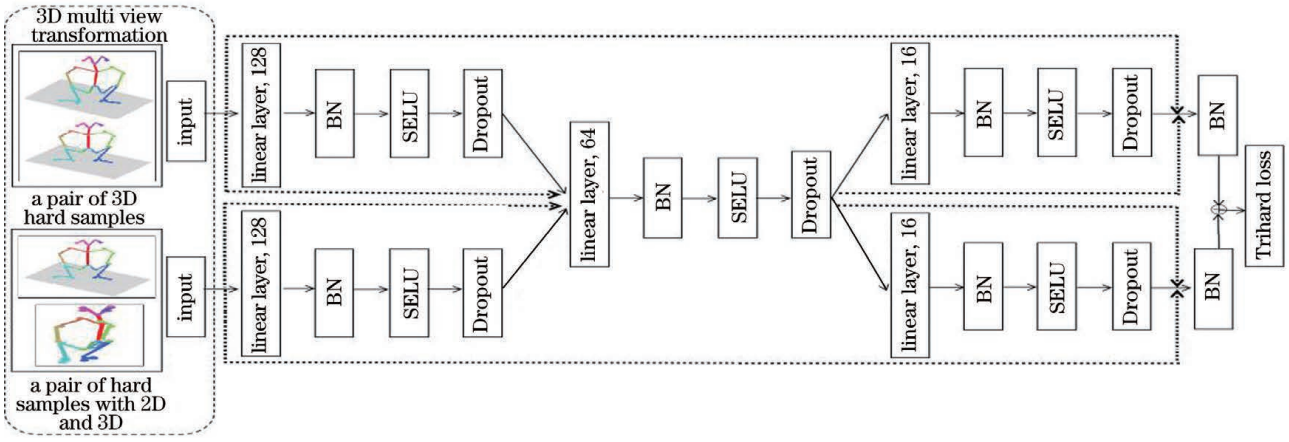


图3　输入为难样本的3D骨架动作识别网络

Fig. 3　3D skeleton action recognition network with difficult input samples

多分支残差网络模型在梯度反向传播过程中容易遗漏流回路径的问题，导致某些学习单元不能获得人体动作特征[27]。因此，提出了一种双通道多特征交叉融合的方法，使两部分网络共用一个残差单元，每个分支只能通过交叉部分网络进行反向传播，解决了遗漏残差单元的问题，也实现了模型的参数交换。每个分支的交叉部分只能对应一个滤波器，同一个滤波器共享两个分支的权重参数，只能提取

一类特征，将这两个分支的 2 组特征进行交叉融合，从而丰富了特征表达，使每个分支的输出特征同时具有两个网络的融合特性。

3　实验结果与分析

3.1　3D 骨架姿态估计的模型训练与评价

用标准数据集 Human3.6M 对本方法中双分支姿态估计网络的表现进行评估。数据集中的 S1、S5、S6、S7、S8 用于训练，S9、S11 用于测试，然后用真实骨架坐标与预测骨架坐标之间的平均误差评估网络的性能，结果如表 1 所示。可以发现，本方法在给出的所有模型中表现最优，至少可将误差减少 2.03%。双分支姿态估计网络可将 2D 姿态转换为 3D 姿态，本网络在北京联众 GPU 集群上进行训练，集群包括四块 GeForce GTX 1080Ti GPU 显卡，显存为 44 GB，该环境下 OpenPose 处理一帧图像的运行时间小于 50 ms。在 Human3.6M 数据集上完成 256 个动作的估计总耗时为 2620 ms，平均每完成一次 3D 姿态估计需要 10 ms，帧频可达到 100 FPS(FPS 为每秒传输帧数)。用两个 3D 姿态估计器分支分别训练相同的参数集，其中，最小的 batch 用 256 个姿态的图像样本，Dropout 为 0.5。初始学习率为 0.002，每 100 个 epoch 误差减少 0.5。1000 个 epoch 后训练结束，可以发现，本方法在上述集群环境下完成训练大约需要 70 min。

表 1　不同方法在 Human3.6M 数据集上的动作估计误差

Table 1　Action estimation errors of different methods on the Human3.6M data set　　unit：mm

Method	Direct	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Avgrage
Ref.[12]	52.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	53.4	67.2	54.8	60.4
Ref.[11]	49.2	51.6	47.6	50.5	51.8	48.5	51.7	61.5	70.9	53.7	60.3	48.9	53.9
Ref.[9]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.3
Ref.[16]	43.2	38.1	40.8	44.4	51.8	43.7	38.4	50.8	52.0	42.1	42.2	44.0	44.3
Ours	44.6	39.5	39.7	41.2	51.1	42.9	40.8	42.9	50.6	40.8	44.6	42.9	43.4

图 4 为本地真实 3D 模型与预测 3D 模型的可视化结果，测试集上的样本输出包括坐姿、跳跃和下蹲的 3D 骨架。对比发现，3D 真实骨架和 3D 预测骨架模型的误差较小，而 OpenPose 采集的 2D 骨架则出现了明显的遮挡问题。

3.2　3D 骨架动作识别模型的训练

实验使用的标准数据集为 NTU RGB+D 60，该骨架视频数据集包含 60 个类别，56880 个样本，每种样本同时包含原始图像及其对应的骨架关键点样本，每个骨架样本包含 25 个骨架关键点及其 3D 坐标。NTU RGB+D 60 数据集的 RGB-D 图像可用于训练和测试 3D 骨架估计的有效性。

在动作识别阶段，用两个 3D 姿态估计器分支分别训练相同的参数集，模型参数用 Xavier 进行初始化，参数优化由 Adam 完成。其中，最小的 batch 使用 10 个姿态的 3D 骨架样本，Dropout 为 0.5。初始学习率设置为 0.001，每 50 个 epoch 减少 0.1，网络在 NTU RGB+D 60 数据集上训练 1500 个 epoch 后结束。本方法在上述集群环境下完成训练大约需要 110 min。在 NTU-RGB+D 60 数据集的每个子集上找到最终的训练权重，该权重参数的大小为 2.5 M。由于本网络的深度较小且结构简单，网

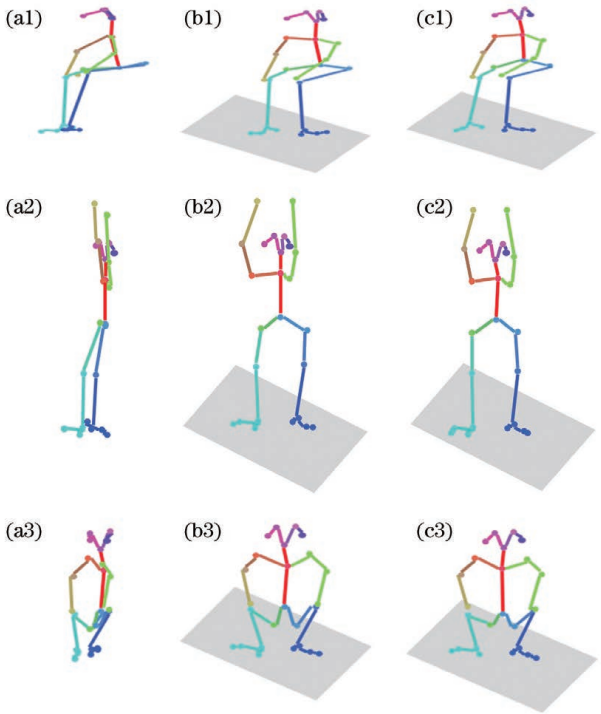


图 4　本地真实 3D 模型和预测模型。(a)OpenPose 采集的 2D 骨架；(b)本地真实的 3D 骨架；(c)估计的 3D 骨架

Fig. 4　Local real 3D model and prediction model. (a) 2D skeleton collected by OpenPose; (b) local real 3D skeleton; (c) estimated 3D skeleton

络预测阶段的计算时间短，一个 batch 的耗时约为 200 ms，识别每个动作所需的平均时间约为 15～25 ms，帧频可达到 50 FPS，可适用于实时动作识别。

3.3 数据增强对识别精度的影响

上述研究表明，融合骨架的 2D 特征和 3D 特征可增强模型的泛化能力，在输入阶段完成视角数据增强对最终测试阶段的识别精度有积极作用。将多视角特征变换数据和未经视角特征变换的数据进行对比实验，除了对比对动作有无旋转扩充的因素外，还对比了扩充数量比例对识别精度的影响，结果如表 2 所示。可以发现，视角扩充的动作数量越多，模型的识别率越高，这表明借助欧拉角旋转公式扩充视角后可提升模型的识别精度，增强模型的视角泛化能力。

表 2 增强数据集后网络的精度

Table 2 Accuracy of the network after enhancing the data set

Expansion ratio	Without treatment	20%	40%	60%	80%	100%
Accuracy /%	82.2	85.1	85.9	86.5	87.9	88.2

为了验证本方法的通用性，用本方法与 2018～2021 年几种最先进 3D 骨架动作识别方法在 NTU-RGB+D 60 数据集上进行对比实验，结果如表 3 所示，评价指标包括交叉受试者(CS)准确率和交叉视野(CV)准确率。可以发现，本方法的 CS 准确率取得了第 4 名，CV 准确率取得了第 2 名，优于多数方法。虽然本方法的CS准确率没有超过第3名的方法，但借助欧拉变换完成数据集扩充后，交叉视野(CV)的 3D 动作识别准确率却排在了第 2 名，这也证明了本方法中多特征融合模型和数据增强的有效性。

表 3 不同方法在 NTU-RGB+D 60 验证数据集上的精度

Table 3 Accuracy of different methods on the NTU-RGB+D 60 verification data set unit: %

Method	Year	CS	CV
RA[28]	2018	85.9	93.5
AS[29]	2019	86.8	94.2
2s-AGCN[30]	2019	88.5	95.1
Two-stream TL-GCN[31]	2020	89.2	95.4
2s-SGCN[32]	2021	90.1	96.2
Ours	2021	**88.2**	**95.6**

3.4 特征融合对识别精度的影响

为了进一步分析骨架的 2D 特征和 3D 特征融合方法对动作识别准确率的影响，通过两方面对比了特征融合的有效性。第一组实验：一个方法在 NTU RGB+D 60 数据集上用 3D 骨架估计作为动作网络的输入，另一个方法用 NTU RGB+D 60 数据集直接给出的 3D 骨架关键点作为输入，以验证本方法通过 3D 骨架估计生成 RGB-D 图像的有效性。第二组实验：对比骨架 2D 特征和 3D 特征融合和只进行 3D 特征识别验证融合网络的有效性，实验结果如表 4 所示。可以发现，通过融合骨架 2D 特征和 3D 特征提取的 3D 骨架估计能更真实地反映 3D 骨架坐标的真实值，且骨架的 2D 特征和 3D 特征融合训练在动作识别阶段表现出了较高的识别率。

表 4 NTU RGB+D 60 数据集上动作估计和多特征融合对识别精度的影响

Table 4 Effect of motion estimation and multi feature fusion on recognition accuracy on NTU RGB + D 60 data set unit: %

CS	After preprocessing of motion estimation network	Using NTU RGB + D 60
With 2D feature fusion	88.2	86.2
Without 2D feature fusion	82.5	81.1

4 结 论

通过实验验证了双分支姿态估计网络将 2D 姿态转换为 3D 姿态的能力，结果表明，相比其他先进方法，本方法至少可将误差减少 2.03%。借助欧拉角旋转公式对视角扩充的模型识别率比无视角扩充的模型识别率高，表明动作识别数据集的视角扩充有利于增强模型的视角泛化能力。融合骨架的 2D 特征与 3D 特征可提高分类器的泛化能力，增强骨架在不同维度的特征互补关系并提升 3D 动作识别的准确率。但在 3D 骨架动作识别中存在一些不足，如不能适应户外可见光环境、只在特定的数据集上有效、用监控实时数据进行测试时识别率较低、在骨架实时动作识别中存在轻微抖动不连续的问题。

此外，当视频中同时出现过多的行人时，3D 骨架动作识别会出现延迟问题，因此，后续还需对这些不足进行进一步研究。

参 考 文 献

[1] Zeng S, Geng G H, Zou L B, et al. Real spatial terrain reconstruction of first person point-of-view sketches [J]. Optics and Precision Engineering, 2020, 28(8): 1861-1871.
曾升, 耿国华, 邹林波, 等. 第一人称视角地形轮廓草图的真实空间重建[J]. 光学 精密工程, 2020, 28(8): 1861-1871.

[2] Zhou M Q, Fan Y C, Geng G H. A spatial symmetry descriptor for 3D model[J]. Acta Electronica Sinica, 2010, 38(4): 853-859.
周明全, 樊亚春, 耿国华. 一种基于空间对称变换的三维模型形状描述方法[J]. 电子学报, 2010, 38(4): 853-859.

[3] Li M W, Shi H Q. Multispectral palmprint fusion recognition based on local joint edge and orientation patterns[J]. Journal of Changchun Normal University, 2020, 39(6): 69-80.
李梦雯, 施汉琴. 基于局部联合边缘和方向模式的多光谱掌纹融合识别[J]. 长春师范大学学报, 2020, 39(6): 69-80.

[4] Jiang C, Geng Z X, Lou B, et al. Automatic image stitching based on scale invariant feature[J]. Remote Sensing Information, 2013, 28(3): 20-25.
姜超, 耿则勋, 娄博, 等. 基于尺度不变特征的影像自动拼接[J]. 遥感信息, 2013, 28(3): 20-25.

[5] Li Y M, Su L. A grid map merging approach based on local feature [J]. Computer Applications and Software, 2020, 37(1): 110-115.
李雅梅, 苏龙. 一种基于局部特征的栅格地图拼接方法[J]. 计算机应用与软件, 2020, 37(1): 110-115.

[6] Huang J J, Ding Y J. Improved video optical flow field estimation based on wavelet transform and HSI [J]. Modern Electronics Technique, 2010, 33(12): 117-120.
黄金杰, 丁艳军. 一种基于小波变换和 HSI 的改进视频光流场估计算法[J]. 现代电子技术, 2010, 33(12): 117-120.

[7] Xiong J L, Wang C. Simultaneous localization and mapping based on RGB-D images with filter processing and pose optimization [J]. Journal of University of Science and Technology of China, 2017, 47(8): 665-673.
熊军林, 王婵. 基于 RGB-D 图像的具有滤波处理和位姿优化的同时定位与建图[J]. 中国科学技术大学学报, 2017, 47(8): 665-673.

[8] Chen C H, Ramanan D. 3D human pose estimation= 2D pose estimation + matching [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5759-5767.

[9] Martinez J, Hossain R, Romero J, et al. A simple yet effective baseline for 3D human pose estimation [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2659-2668.

[10] Moreno-Noguer F. 3D human pose estimation from a single image via distance matrix regression[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1561-1570.

[11] Mehta D, Sridhar S, Sotnychenko O, et al. VNect: real-time 3D human pose estimation with a single RGB camera [J]. ACM Transactions on Graphics, 2017, 36(4): 44.

[12] Sun X, Shang J X, Liang S, et al. Compositional human pose regression[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2621-2630.

[13] Yang J, Zhang S J, Zhang C H, et al. Research on human action recognition and contrast based on OpenPose[J]. Transducer and Microsystem Technologies, 2021, 40(1): 5-8.
杨君, 张素君, 张创豪, 等. 基于 OpenPose 的人体动作识别对比研究[J]. 传感器与微系统, 2021, 40(1): 5-8.

[14] Cippitelli E, Gasparrini S, Gambi E, et al. A human activity recognition system using skeleton data from RGBD sensors [J]. Computational Intelligence and Neuroscience, 2016, 2016: 4351435.

[15] Aubry S, Laraba S, Tilmanne J, et al. Action recognition based on 2D skeletons extracted from RGB videos[J]. MATEC Web of Conferences, 2019, 277: 02034.

[16] Pham H H, Salmane H, Khoudour L, et al. A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera[J]. Sensors, 2020, 20(7): 1825.

[17] Yang F, Wu Y, Sakti S, et al. Make skeleton-based action recognition model smaller, faster and better [C] // Proceedings of the ACM Multimedia Asia, December 16-18, 2019, Beijing, China. New York: ACM, 2019: 1-6.

[18] Ionescu C, Papava D, Olaru V, et al. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325-1339.

[19] Zhang H W, Yu Z Z, Lei W W. Study on the basic theory of quaternion and Eulerian equation of the rotating vector[J]. Journal of Geodesy and Geodynamics, 2020, 40(5): 502-506.
张捍卫, 喻铮铮, 雷伟伟. 四元数的基本概念与向量旋转的欧拉公式[J]. 大地测量与地球动力学, 2020, 40(5): 502-506.

[20] Shahroudy A, Liu J, Ng T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1010-1019.

[21] Sigal L, Balan A O, Black M J. HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International Journal of Computer Vision, 2009, 87(1/2): 4-27.

[22] Luvizon D C, Picard D, Tabia H. 2D/3D pose estimation and action recognition using multitask deep learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5137-5146.

[23] Guo F Z, Kong J, Jiang M. Action recognition based on adaptive fusion of RGB and skeleton features[J]. Laser & Optoelectronics Progress, 2020, 57(20): 201506.
郭伏正, 孔军, 蒋敏. 自适应融合 RGB 和骨骼特征的行为识别[J]. 激光与光电子学进展, 2020, 57(20): 201506.

[24] Liu F, Yu F Q. Human action recognition based on global and local features[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021004.
刘帆, 于凤芹. 基于全局和局部特征的人体行为识别[J]. 激光与光电子学进展, 2020, 57(2): 021004.

[25] Hou C P, Jiang T L, Lang Y, et al. Human activity and identity multi-task recognition based on convolutional neural network using Doppler radar[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021009.
侯春萍, 蒋天丽, 郎玥, 等. 基于卷积神经网络的雷达人体动作与身份多任务识别[J]. 激光与光电子学进展, 2020, 57(2): 021009.

[26] Huang Y W, Wang F, Li J H, et al. Algorithm for video temporal action proposal combining watershed and regression networks[J]. Chinese Journal of Lasers, 2019, 46(11): 1109001.
黄韵文, 王斐, 李景宏, 等. 结合分水岭和回归网络的视频时序动作选举算法[J]. 中国激光, 2019, 46(11): 1109001.

[27] Li Y, Yang D D, Han Y J, et al. Siamese neural network object tracking with distractor-aware model[J]. Acta Optica Sinica, 2020, 40(4): 0415002.
李勇, 杨德东, 韩亚君, 等. 融合扰动感知模型的孪生神经网络目标跟踪[J]. 光学学报, 2020, 40(4): 0415002.

[28] Song Y F, Zhang Z, Wang L. Richly activated graph convolutional network for action recognition with incomplete skeletons[C]//2019 IEEE International Conference on Image Processing (ICIP), September 22-25, 2019, Taipei, China. New York: IEEE Press, 2019: 1-5.

[29] Li M S, Chen S H, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3590-3598.

[30] Shi L, Zhang Y F, Cheng J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, Long Beach, CA, USA. New York: IEEE Press, 2019: 12018-12027.

[31] Zhu G M, Zhang L, Li H S, et al. Topology-learnable graph convolution for skeleton-based action recognition[J]. Pattern Recognition Letters, 2020, 135: 286-292.

[32] Yang W J, Zhang J L, Cai J J, et al. Shallow graph convolutional network for skeleton-based action recognition[J]. Sensors, 2021, 21(2): 452.