参考英文文献：<https://link.springer.com/article/10.1007/s11760-020-01644-0>

## 标题 基于骨架序列用于行为识别的旋转时序特征

## 摘要

最近，基于骨架的行为识别方法凭借将从骨架序列中获取到的不同特征作为分类深度模型的输入已经取得了巨大的进步。相比于其他直接使用节点位置信息作为分类任务的表示特征的论文，我们的论文使用了嵌 入在骨架中的相关关系，以减轻因视角多样性引起的影响。特别地，我们提出了一种新型的特征描述量，它基于骨架中地旋转关系以表达一个特定的行为。几何代数（GA）根据GA中称为转子的特定操作，以计算和推导骨架中的旋转联系。为了充分利用一个骨架序列中的空间和时序的特征，我们还设计了两个不同的基于转子的特征描述量，它们分别基于骨架的一帧和相邻两帧的两个骨架的信息。其次，我们提出了一种高效的特征编码策略，以将每一种特征描述量转化为一张RGB图片。紧接着，我们提出了一种基于双流CNN模型的框架，以学习生成于每隔骨架序列的RGB图像，之后又会将两个网络模型的输出进行融合以得到最终的识别准确率。基于NTU RGB+D，Northwestern-UCLA，Gaming 3D，SYSU和UTD-MHAD数据集的广泛实验结果也证明了我们方法的优越性。

**关键词：**行为识别 骨架序列 旋转特征 时空特征 几何代数

## 1 引言

在早期的人体行为识别工作中，传统方法大体上是基于RGB数据进行行为识别。相较于RGB数据，骨架数据不仅仅对北京噪声和光亮变化不敏感，而且对人类行为的高级特征也不敏感。因此，目前越来越多的基于骨架的行为识别方法被提出。

最近，很多研究者主要使用了CNN和RNN模型来获取骨架数据的深层特征。基于CNN的方法能非常的高效地取得不错的结果，这是因为它强大的获取相邻帧和长时间序列相关联系的能力。然而，大多数的努力局限于将节点的特征作为CNN模型的输入，这种做法通常会在对骨架数据编码的过程中，丢失很多重要的结构信息。

对于基于骨架数据的行为识别方法，来自不同视角的骨架数据对行为识别造成了一定的挑战，因此，直接将节点的绝对坐标作为特定行为的特征表示，可能会引起训练样本和测试样本中视角的不匹配问题，这个问题将会降低识别的准确率。因此，我们的工作充分的发挥了骨骼序列中的相对几何关系，尤其是骨骼结构中两个骨边的旋转关系。其次，基于旋转的时空特征的学习方法能够充分利用到嵌入在骨骼序列中的空间配置信息和动态特征。在我们的论文中，由于几何代数空间对子空间和向量的几何运算简单而且有力，因此我们几何代数空间中计算和推导这种旋转关系。特别地，GA空间中的旋转操作的转子被用来计算骨骼中的旋转联系。然后三个角度量会被推导且被定义为基于转子的特征描述量。我们的文章提出了基于转子的针对每个行为的空间和时序特征描述量，以便于充分利用单帧骨骼和相邻帧骨骼中的旋转关系。除此之外，我们提出了一种高效的特征编码方法，它能够将特征描述量转化为RGB图片。最终，我们设计了双流CNN模型，一个是用来学习基于旋转的空间特征描述量，而另一个则用来学习基于旋转的时序特征描述量。由此，我们的双流CNN模型就实现了对一个骨骼序列的时空特征的建模。我们还将每一个分支的分类结果融合以得到最终的来识别准确率。

我们工作的流程图大致如图1所示, 总而言之，我们的贡献可以总结为以下几点：

(1) 提出了一个新型特征描述量，它基于几何代数计算出的骨骼中的旋转关系。

(2) 提出了两个基于转子的特征描述量，以充分挖掘骨骼序列中的空间和时序的特征。

(3) 在四个具有挑战性的数据集上的仿真实验表明了我们方法的有效性。

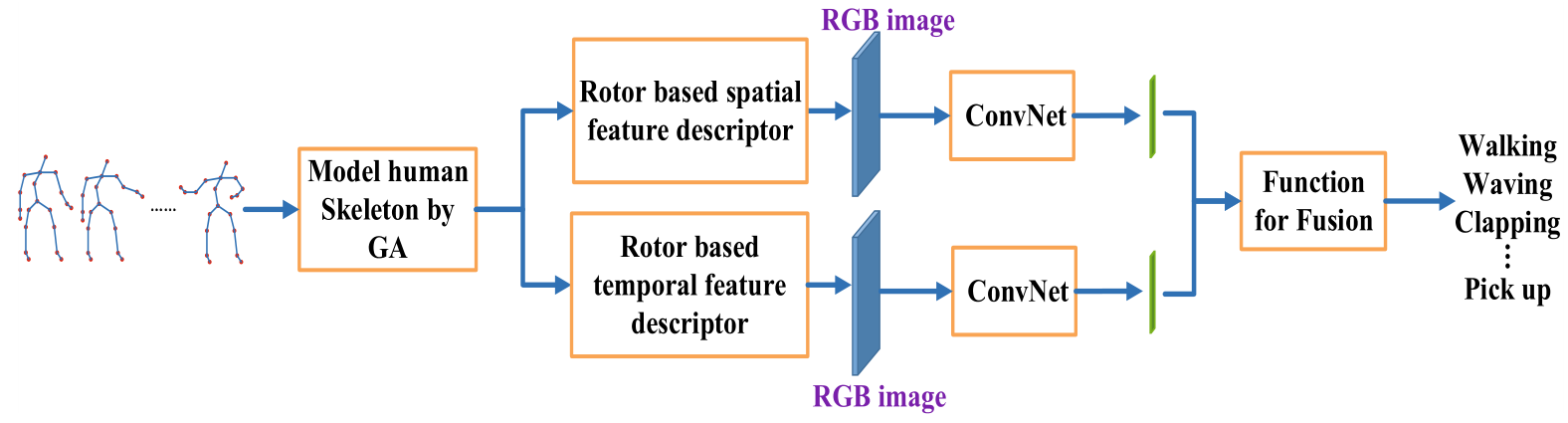


图 1 所提出模型的流程图

## 2 相关工作

在这个部分，我们简要地回顾现存的工作，它们于我们所提出的方法紧密相关。在基于骨骼的动作识别方法方面的相关工作中，研究者们已经应用了这种高效的推导自行为识别的骨骼数据中的几何特征信息。Ferd[10]等人将两个相邻骨骼的角度作为特征，以选择最具有信息量的节点用于分类任务。Vemulapall[1]等人将两个身体部分的旋转关系作为Lie Group中的几何特征。Huang[2]等人扩展了目前的工作，并且将Lie Group的结构融入到深度模型中，以学习基于旋转信息的特征。Zhang[8,9]等人探索并总结了不同类别的几何特征，比如位置、角度和两个不同身体部分的距离，以对比这些特征的性能。特别地，Li[11]等人利用几何代数提出了一种基于骨骼序列地视角不变性运动的表示方法。最近，越来越多的文献利用CNN模型来学习骨骼特征并且取得了很好的性能。Ke[3]等人将每个骨骼序列转换成三个clip片段，同时使用CNN模型来学习时空特征。Liu[4]等人使用CNN模型来提取来源于骨骼数据的色彩图中具有鲁棒性和鉴别性的特征。除此之外，Hou[12]等人使用了一个标准的卷积网络的架构来学习骨骼序列中具有鉴别性的特征。Tang[5]等人使用基于图结构的CNN模型捕获节点之间的依赖性，并且和其他先进的方法相比，取得了非常有竞争力的结果。

## 3 所提出的模型

我们所提出的方法由三方面组成，它能够使用骨骼序列中的旋转关系来生成一张RGB图像用于分类任务。根据骨骼数据生成RGB图像的流程图如图2所示。

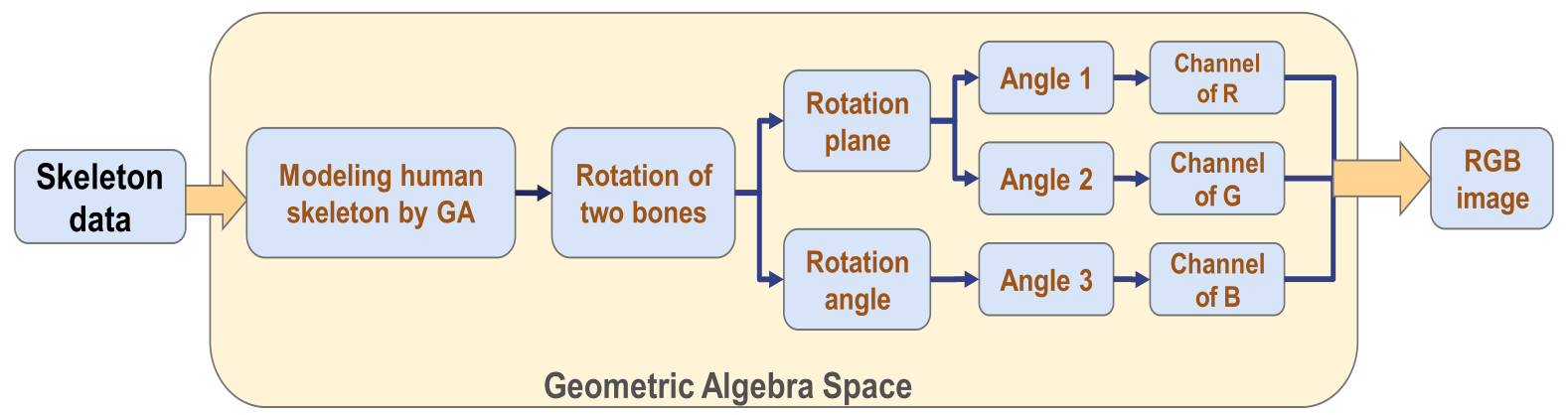


图 2 通过基于转子的特征描述量生成RGB图像的流程图

### 3.1 几何代数空间中对人类骨骼的描述

人类行为可以表达成骨架的运动过程，而每一个骨架是由关节和骨骼组成。这种骨骼的节点可以被视为3D欧几里得空间中的节点，这些节点能够生成**各种几何实体**作为人体骨骼的表示。而众所周知，几何代数非常适合解决哪些直观的几何问题。因此，我们用3D几何代数空间中的几何模型，以描述人类骨骼。

在几何代数中，几何积被引入。对于两个向量和,他们的几何积可以被表达成他们内积和外积的加。

(1)

内积即是点乘，产生一个标量，然而外积会得到一个新的量，称之为二维向量，它代表了一个有方向和有限长度的平面。由于节点的坐标是处于3D的欧几里得空间中，所以我们定义3D的几何代数空间为，拥有一个正交的向量基，表现为，这个正交的向量积通过不同的集合可以表现标量、向量、二维向量和三维向量。

（2）

(3)

接着，我们对于一个给定的有帧的骨骼序列，用来表达第帧的第个骨骼节点。的3D坐标表达成，其中表示节点的总数，而几何代数空间中的可以被一组正交基表达。

(4)

骨骼，连接了两个相邻节点和，且其可描述成如下的向量：

(5)

基于以上的定义，嵌入在每帧的骨骼数据中的，用于表达相关空间信息的描述量将在下面的部分得到推导。

### 3.2 基于转子的特征描述量的推导

首先，我们呈现旋转表达式的推导过程，接着，提出两类基于转子的特征描述量。在几何代数空间中，任何的几何实体，旋转信息都可以通过转子来表达，而旋转信息可以接着被定义为：

(6)

其中，是的逆阵形式，而在几何代数中可以被表达成：

(7)

在这里，表示几何代数中的一个二维向量，表达的是旋转平面，而表示的是一个标量，表达旋转的角度。

在本论文中，我们根据公式(2)表达二维向量。

(8)

是标量，由于且是一个的二维向量。由于标量的约束条件，我们给出以下定义，以得到的唯一解。

(9)

其中。因此我们用两个角度来表达旋转平面。最后，转子被表达为：

(10)

很容易地可以推导出，转子的具体公式，完全由三个角度**确定**。

(11)

在本文中，我们使用这三个角度作为基于转子的特征描述量，以表示旋转的几何关系。

#### 3.2.1 基于转子的空间特征描述量

在这篇文章中，我们用一帧中的两个骨骼之间的旋转关系来表达骨骼结构中相关的空间信息。旋转西悉尼可以用几何代数中的转子来计算。我们定义第帧中的两个不同骨骼为和。通过两个相邻节点连接，亦然，如图3所示。我们使用同一帧中两个不同骨骼的旋转关系，以表达基于转子的空间特征描述量。

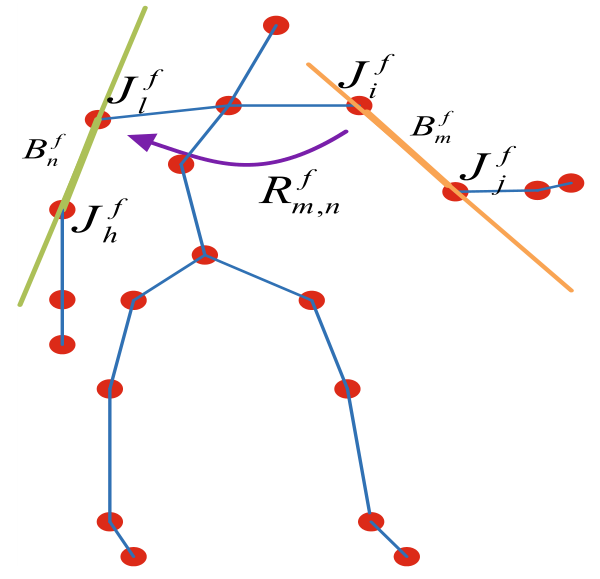


图 3 骨骼和的旋度

3D几何代数空间中，和的公式同样可以用公式(5)以及相应的邻近节点的3D坐标表达。如果仔细观察以上转子的推导公式，我们会注意到用于旋转的几何实体的范式对推导转子并不必要。因此，不失一般性，我们假定和的范式等于1，以降低计算的复杂性。

我们使用转子来表达和的旋转关系，正如以上推导的那样，由旋转角度和旋转平面组成，这一点参考公式(7)。

首先，旋转角度可以借由和的内积推导得到：

(12)

除此之外，和两个向量全处于旋转平面中，以至于可以被和的外积决定。因此，和两个向量构成了一个二维向量，以表达旋转平面：

(13)

根据所涉及节点的3D节点坐标，我们可以进一步地计算出的具体公式。

(14)

*,*

（15）

相较于公式（8）、（9）和（14），我们可以直接计算两个角度的值，而这两个角度的值直接决定了旋转平面。

最后，我们使用三个角度来构成基于转子的空间特征描述量，表示任意两个骨骼的旋转信息。

(16)

#### 3.2.2 基于转子的时序特征描述量

为了描述几何相关信息不仅仅存在于单帧的一个骨架中，我们提出另一个描述量，它通过相邻帧中的骨骼的旋转信息来表达基于转子的时序特征描述量。

特别地，这第二个特征描述量是由相邻两帧中各自的两个骨骼间的转子计算得到。表示第帧中连接两个节点的骨骼。其中表示第帧中连接两个节点的骨骼。值得注意的是，和被视为几何代数中的向量。正如图4所示，转子用来表达两个向量之间的旋转关系，同时它被计算以推导出时序特征描述量。

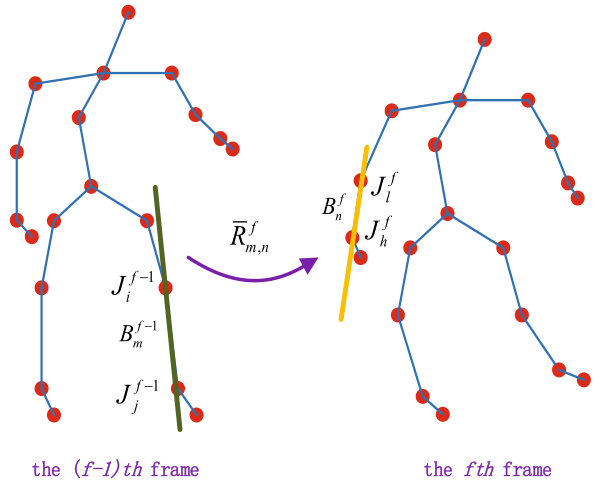


图 4 相邻两帧中和的旋度

时序特征描述量的推导过程于3.2.1节内容相似。三个角度分别表示为，它们构成了，参考公式（8）、（9）（14）和（15），我们通过相关节点的三维坐标可以得到三个角度的值。

(17)

### 3.3 生成RGB图片的编码方法

在这个部分，我们将所有帧以及每个骨骼序列中的所有骨骼通过将其编码成一张RGB图片，以充分地表达一个特定的行为。

图5显示了我们所提出的编码方法的生成RGB图片的结构。

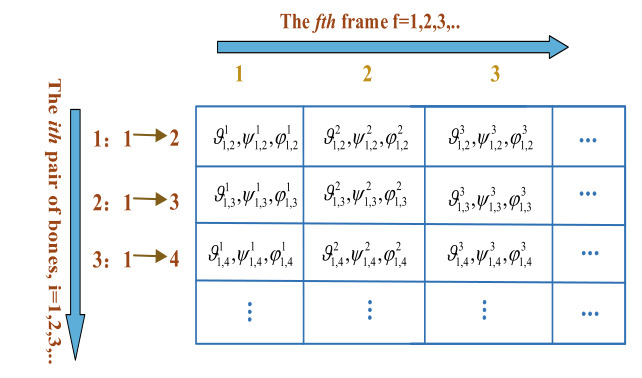


图 5 所生成的RGB图片的结构

图片的列对应了骨骼序列中的第帧，而行对应了第对骨骼，且该两个骨骼分别用和来表达。除此之外，这三个元素于对应的色彩图片的每一个像素中的一一对应。

因此，位于位置的色彩值可以表达为：

(18)

鉴于3D骨骼和图片中的不同值，我们将像素值归一化到0-255，并且为了方便，每个图片归一化到224224后作为深度网络的输入。正如上面提到的，两个特征描述量都基于同一个骨骼序列。因此，对于每一个行为都有两个不同的色彩图片生成，同时被送入分类的深度模型中。在不同的深度网络中，CNN凭借其强大的表达能力，被普遍应用于基于骨架的行为识别方法中。在我们这篇文章中，我们涉及了一个双流的基于CNN的网络模型，以学习来自两个色彩图片的深度特征。

另外，为了探索每个CNN模型生成的深度特征的互补性，我们通过加权以融合来自两个分支的结果。具体来说，我们将两张照片作为两个分支的输入内容，然后，用表现由图片集所在的分支的输出获得的后验概率。它同时表达的是图片属于第个行为类别的可能性。与It图片集相似的是，后验概率表示为，因此，对于一个骨骼序列，它的类别分数是由加权的融合方法推导得出的：

(19)

而和 是两个分支的预测值的相应权重，并且两个值的总和等于1。最后，该序列的标签经由它的类别分数值得出。

## 4 实验

我们在五个常用的基准数据集上对该方法进行了评价和实证分析。数据集中骨骼的节点坐标值已经被给定。以下的五个数据集是NTU RGB+D，Northwestern-UCLA，Gaming 3D，SYSU-3D和UTD-MHAD。接下来将会展示实验的细节和结果。

### 4.1 实验细节

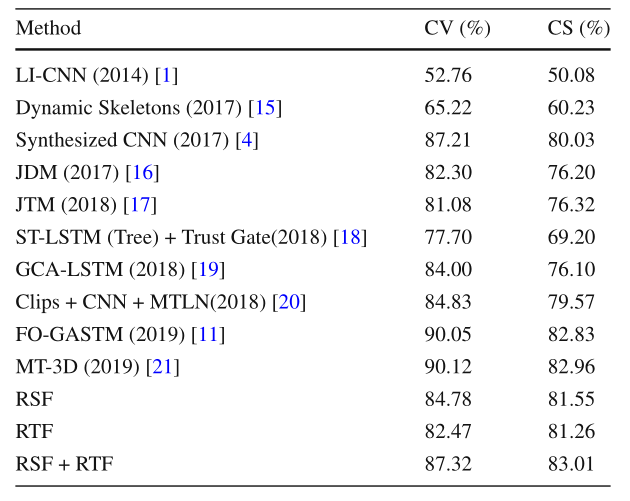
双流CNN模型用于获得最终的预测结果。其中，基于时空特征的图片，分别作为每个网络的输入。我们使用基于ImageNet分类数据集而预训练好的ResNet-50模型。Batch的大小在NTU RGB+D数据集中设置为32，而其他数据集中设定为16。实验执行基于Pytorch框架，使用了两块英伟达GTX 1080 GPU。首先我们将实验中的两个权重值设定为0.5。网络模型的权重借助Mini-Batch 随机梯度下降进行微调，优化器的动量值为0.9，权重下降率为0.0005，学习率设定为0.001，最大的训练循环是100次，同时学习率每20个循环下降一次。

### 4.2 NTU RGB+D 数据集

该数据集是目前对打的骨骼数据集之一，它包含56880个视频样本用于人类行为识别。

它包含60个不同的行为类别和40个人物对象，而每个人物对象，包含25个节点，每个行为经由三台处于相同高度但不同水平角度的照相机捕获，该数据集的两个基准是：(1) Cross-subject（CS）：来自20个人物对象的40320个样本用于训练，其他的样本用于测试。(2) Cross-View（CV）：由照相机2和3捕获得到的样本可以用于训练，而相机1的样本用于测试。

表1 NTU数据集的对比结果



在NTU数据集中的结果和对比实验结果如表1所示，我们所提出的基于转子的空间特征和时序特征的在CV中准确率分别达到84%和82%，除此之外，我们的方法在特征融合后可以达到87%和83%，这一点表现得比表1列出的现在最先进的技术的大多数要优越。

结果也证明了我们的方法的高效性，值得注意的是，我们的方法相较于其他基于几何特征的工作取得了意义重大的提升。不仅如此，我们的方法比一些端到端的可训练的深度模型要表现优异。除此之外，我们可以观察到[11]和[21]论中的工作达到了比我们方法更好的性能。

主要的原因如下示：[21]论文考虑了节点的不同几何特征，包括位置和不同视角下的相对关系，然而我们的方法仅仅使用了一种几何特征即两骨骼间的旋转关系来表达骨骼。[11]论文使用了一个四分支CNN网络框架以学习不同的预定义的活跃的表示信息，同时数据预处理过程实现了对视角的不变性。然而，在我们的方法中仅仅使用两个分支的CNN模型来获得最终的结果，同时我们寻求嵌入在人类骨骼数据中的相关旋转信息，以处理视角多样性的问题而不是使用数据预处理的方法。

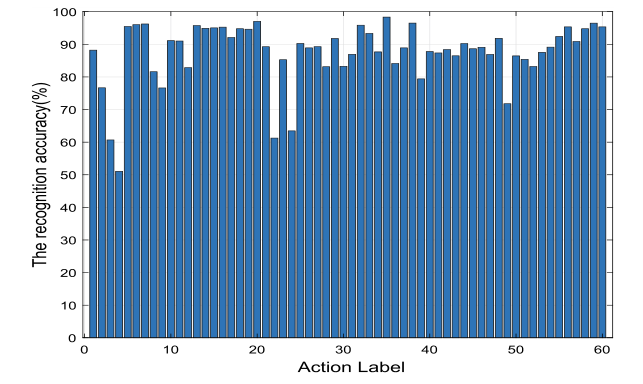


图6 CV下的NTU数据集中每个行为类别的识别准确率

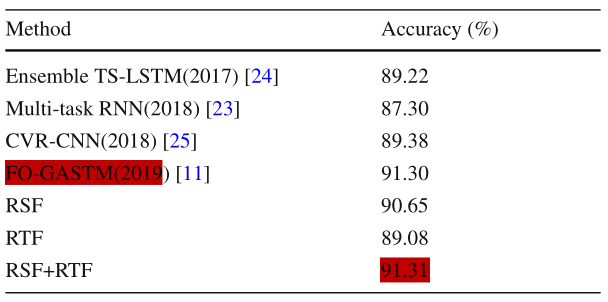
图6显示了在CV情况下，每个类别的识别准确率的分布直方图。水平轴的索引表示行为的id值，如[14]论文中表现得一致。值得注意的是，对于大多数的类别，我们的方法表现得不错，并且取得了大约90%识别准确率的结果。然而，少数的行为比如梳理头发的4号行为，其识别效果并不好，这是因为这些手势行为与其他的行为很相似。

### 4.3 Northwestern-UCLA 数据集

该数据集包含10个行为类别的1494个骨骼序列。它包含了由10个不同人物表现的10个行为，每一个行为都是在三个不同视角下进行。每个骨骼人物都包含了20个节点。我们遵循22论文中的评估协议，使用前两个相机的样本作为训练数据，而第三个相机的样本作为测试数据。

在该数据集上的对比试验结果如表2所示，我们的方法取得了有竞争力的结果，识别准确率高达91.31%，这一点要超过目前最先进的方法。我们方法的表现要由于基于BNN和LSTM模型的结果。除此之外，我们方法的结果略微高于四分支CNN模型的结果，这也表明了我们的模型对不同的视角具有一定的不变性。

表2 N-UCLA数据集的对比结果



我们的方法在该数据集中表现的混淆矩阵如图7所示。通过仔细检查该矩阵，我们可以发现大多数的行为被显著区分了，这一点就证明了我们所提出的方法的高效性。同时我们也观察到意义重大的分类错误发生在举起双手和搬运两个类别中。主要的原因，是这些行为包含了与物体的交互行为，这就使得仅通过骨骼数据无法简单地区分。

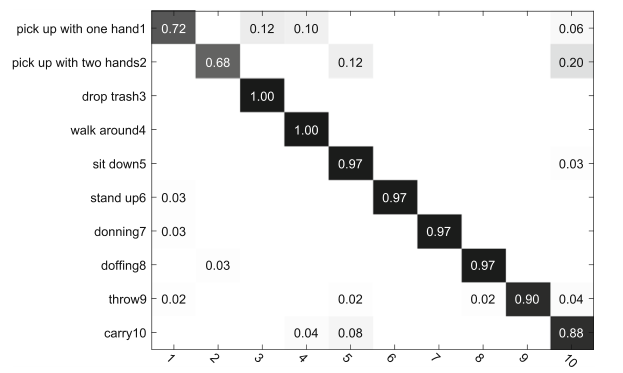


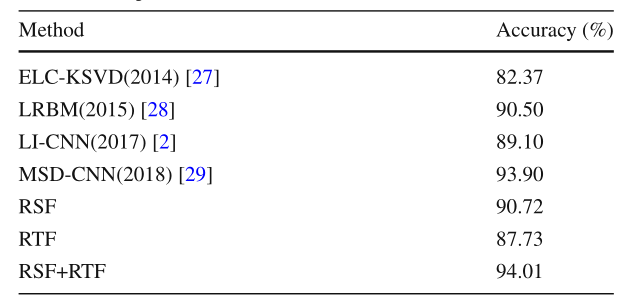
图7 N-UCLA数据集的混淆矩阵

### 4.4 Gaming 3D 数据集

该数据集包含了10个人物展现的20个游戏行为，包含总共663个骨骼序列。我们使用来自该数据集中的骨骼数据，这些数据根据不同的人物进行划分，其中前五个人物集用来训练，剩下的5个人物集用来测试。

G3D数据集的结果和对比实验由表3展现。值得注意的是，我们的方法经过特征融合后达到了94.01%的是被准确率，比其他的模型都要好。这样的结果表明我们的方法在小规模数据集中表现依旧很好。

表3 G3D数据集的对比实验



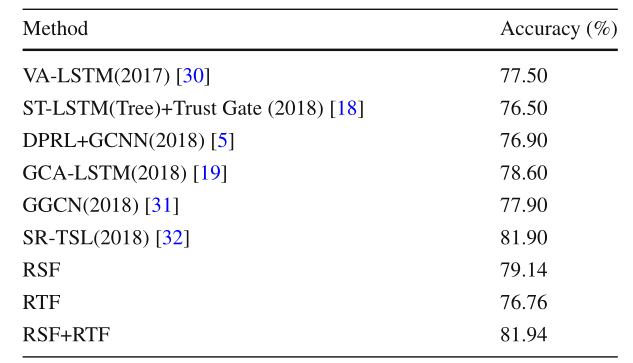
### 4.5 SYSU-3D数据集

该数据集重点关注人类和物体的交互行为，40个参与者展现12个不同的行为，包含了480个视频片段。我们使用前20个人物集的样本用于训练，而其他的20个人物集用于测试。我们采用30折交叉验证的方式。

该数据集的实验结果与对比试验如表4所示。RSF、RTF和两者融合的平均准确率分别做了展示。值得注意的是，我们所提出的方法表现比最先进的技术要好。正如先前解释的，我们模型中的新型的特征表示能够获取到嵌入在骨骼中的相对几何关系，而绝对的节点坐标并不重要。

进而，相对的几何关系可以直接表现人类的特定姿势，因为这些姿势在人们和物体互动的额时候并不相同。这可能也就是我们的模型表现比其他模型好的原因所在。

表4 SYSU数据集的对比结果

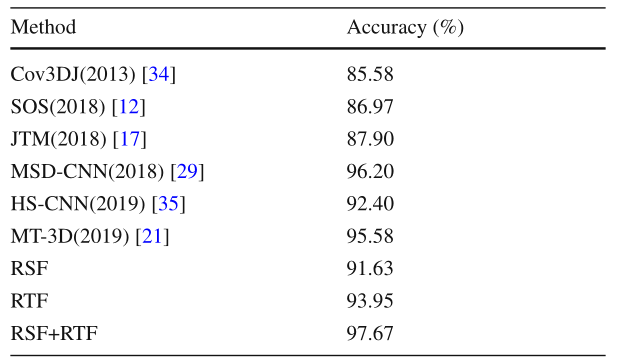


### 4.6 UTD-MHAD 数据集

该数据集是一个多模态动作数据集，由微软的骨骼相机和一个可穿戴性传感器采集而成。它包含由8个人物（4男4女）展现的27个行为，并且每个人物展现每个动作4次。在一处了三个中断的片段后，这个数据集大概有861个序列。对于该数据集而言，1.3.5.7号人物的样本用于训练，而其他的用于测试。

我们的方法与其他最先进技术的表现对比结果如表5所示。显然地，我们的方法在该数据集中表现得比其他方法要优越。

表5 UTD-MHAD数据集的对比实验



## 5 结论

在这篇文章中，我们研究了时空特征对骨骼序列的旋转信息学习的高效性。我们也描述了人类骨骼同时推导了在几何代数空间中相邻骨骼的旋转信息的推导过程，以生成我们所提出的特征描述量。更具基于转子的特征描述量的推导过程，我们的方法有效地利用了几何代数中简单表示和简单计算的优点。除此之外，我们提出的特征编码策略能够简单地将特征描述量转化为RGB图像，以作为CNN分类模型的输入。更进一步地，双流模式对增强模型性能很有帮助。在四个具有挑战性地基准数据集上进行地实验结果，证明了我们方法地有效性。而对于未来，我们也将拓展我们的框架，以解决多人参与的行为问题。