## 收获：

1 对比实验应该使用各自最优的参数组合！

## 基于Kinect相机的行为识别方法的对比综述

## 摘要

基于视频的人类行为识别的方法是当前计算机视觉中最火热的研究领域之一。

不同的研究均表明了是识别的性能非常依赖于提取特征的类型以及对人类行为展现represent的方式。

Kinect相机的出现，引起了学术上一系列基于骨架的人类行为的识别方法，然而，**在特征类型的分组下**，目前缺少对这一系列方法的综合对比，比如手工制作的特征与深度学习特征以及基于深度depth和基于骨架skeleton的特征。

而在我们这篇paper中，我们选择了10个基于骨架的算法，分析且比较了这些算法都在cross-subject和cross-view两种角度的数据集上的表现。另外，我们在对比过程中实现并改进了其中部分的算法及其变体，同时实验表明大多数的方法在cross-subject类型的数据集上的表现要普遍由于cross-view类型；除此之外还表明**基于骨架的特征**在cross-view类型的数据集中的表现要比**基于深度的特征**的鲁棒性要好；最后实验表明**深度学习的特征**更适合大规模数据集。

## 1 引言

人类行为识别具有很多有用的应用，比如人机交互，smart video surveillance智能视频监控，运动和health care保健。然而尽管学生上出现了大量相关的研究文献，但是需多仍有很多问题影响着这些研究的算法的性能，比如，视角不同性、外观visual appearances、身体大小和照明条件以及动作进行的速度。

除此之外，还有一些问题，比如行为主体被场景中的其他物体部分遮挡，以及行为主体之间的自行遮挡。

有些早期的研究是使用传统的RGB视频，尽管这些基于视频的研究取得了不错的结果，但是它们的识别准确率仍然相对低，即使是在当前场景中没有杂乱的东西，free of clutter。

Kinect相机本来是一个为了**拓宽XBOX 360用户的3D游戏体验，**然而因而该相机能够捕捉实时的**RGB和depth视频**，所以它成为了一个公开可用的工具箱，用于从深度视频的每一帧中**计算人体骨架模型**，许多基于Kinect相机进行行为识别的研究论文已经出现了。

深度视频相比传统RGB视频的好处在于即使场景杂乱，它仍能够容易地分割前景中的人类对象。由于depth视频没有色彩西悉尼，穿着在人体上的衣服的色彩对分割过程没有影响。这就使得研究者可以把精力更多地放在获取健壮的特征描述符**以描述动作而不是低层次的分割任务。**

使用depth的视频的论文有9-14；

但是**depth图像也很容易受到各种因素的影响而产生噪声，见15。**

因此使用depth图片并不能总是保证好的行为识别性能。

论文16提出了一种基于人体骨架跟踪框架的算法用于计算人体骨骼模型的三维关节位置坐标。

这种跟踪框架除了能够提供实时的depth视频流以外，**它还开辟了基于骨架的人体动作识别的研究领域**。

**以上内容介绍了各种人体行为识别的方法的发展过程。**

使用Kinect相机的行为识别方法**根据特征描述feature descriptors被提取的方式**可以被分成两个类别。

第一种是手工制作的特征，相关的研究方法需要两个复杂的手工设计阶段，即特征提取以及特征表达（**区别在哪儿？提出来的东西不就是表达的内容吗？**）以build构建最终的特征描述。需要指出的是，特征提取和特征表达两个阶段的内容因模型不同而不同。

特征提取阶段：该阶段可能涉及计算depth、color的梯度、分布直方图以及其他更多复杂的视频transform的方式。

特征表达阶段：该阶段可能涉及对从上一阶段提取得到的一系列的特征因素的合并concatenation，或者是对这些特征新宿的更为复杂的fusion操作，以获取到最终的特征描述。

以上的这些方法的问题是：通常需要研究者继续宁仔细的特征工程和调参。相关的文献有9-13、16、17、19、23、24。

第二种是深度学习得到的特征。

深度学习的方法降低了对特征工程的需要；然而这些深度学习方法需要大量的标记的训练数据和相当长的训练时间，其中对训练数据的高要求使得数据不容易得到。

而这种方法的缺点是，相比之下，对于小规模数据集，深度学习方法可能不能够获得最好的性能。

相关的文献有：14、18、20-22、25-38。

### 研究贡献

尽管手工制作的特征和深度学习得到的特征被用于行为识别，但是据我们所知，目前对于两类方法的综合对比目前还没有。因此我们paper的贡献有两个方面：

1 评估了10种最新的算法的性能，特别侧重于比较使用手工特征与深度学习特征以及基于骨架特征和基于depth的特征的有效性。

2 评估了这些算法的跨视角和跨行为主体的性能。而对于多视角数据集，我们还评估了相机视角对于大小规模数据集的影响，即使所使用的特征是基于深度的，基于骨架的还是基于深度+骨架的。

## 2 相关工作

行为识别方法有三类：

1 基于色彩；1-8、39-55。

2 基于depth；9-14、22、24、25、47、55-65

3 基于骨架；16-24、28-38、55、66-76。

### A 基于depth的行为识别

Depth的识别方法越来越流行，是因为它提供了实时的具有成本效益的传感器（because of the availability of real-time cost-effective sensors）。

现有的基于depth的动作识别方法大多采用全局特征，比如**时空volume和information**。

比如，论文11将depth序列中获得的4D曲面法线投影到4D规则空间中，以捕捉到**识别特征（discriminative features）**，且构建了HON4D Histogram of Oriented 4D Normals。

论文60扩展了extendHON4D，将depth视频中的局部相邻的超曲面发现连接起来，共同刻画局部形状和运动信息。具体来讲，**他们引入了一个自适应的时空金字塔，将depth视频细分为一组时空单元，以获得更具有辨别力的特征。**

论文57提出了从depth传感器中过滤掉噪声，从而得到更加可靠的时空兴趣点，用于行为识别。

尽管这些方法在**正面动作识别frontal action recognition方面取得了令人深刻的性能，**但是它们**对视角的变化**非常敏感。而缓解视角问题的方法之一是直接处理点云pointclounds，见论文13。

以上方法均是手工制作的特征的方法，接下来是深度学习得到的特征的研究。

论文61，使用层级depth运动图**以提取身体形状和运动信息，再在HDMM基础上训练了三通道的CNN模型。**

论文14也提出从~~真实运动捕获数据中~~训练**单个个体的姿势模型**HPM模型，以将**不同未知视图中的人体姿势**转换到**视角不变的特征空间**。

论文64使用双流模型同时**学习动作属性之间的语义关系**。

### B 基于骨架的行为识别

现在基于骨架的行为识别的方法分成两类：基于节点和基于身体部位。

#### 1基于节点的方法

使用从OpenNI框架中提取到的节点的坐标对节点的位置和运动进行建模。

比如，一个参考节点被设定后，其他节点的坐标将根据该参考节点被定义；或者节点的方将会相对于一个固定的坐标体系进行计算，同时同来表示人类的姿势。66号论文

#### 2对于基于身体部位的方法。

在该方法类别中，身体部位将用来构建人类的关节系统；通常的身体部位被建模为通过节点连接的固定的圆柱体。

**类似节点角度、身体部位的时序演变和身体部位之间的几何关系这类信息都被用来表示人体姿势**。**目前还有旋度特征用来表达姿势。**

论文17使用一对身体部位之间的几何关系，作为一个点……

论文20依赖身体部位和身体节点两方面的信息，人体骨架模型被分成5个部分，同时从每个部分中选择一个特殊的节点作为参考节点，而其他的节点将根据该点表示成vectors。不仅如此，**多种距离衡量指标将根据这些向量进行计算，其值构成每个视频帧的feature vector，而所有视频帧中的特征向量将最终被添加在一起，并归一构成一个手工制作的灰度图以送入CNN模型**进行运算。

与论文20相似的一种方式是论文77提出的，使用由身体节点形成的核以得到特征图，从而将特征图喂入CNN模型以同时进行行为识别和领域适应domain adaptation。

最近的人类行为识别的文章倾向于favor深度学习技术以进行行为识别。除了上面提及的CNN模型，还有循环神经网络。这是因为LSTM可以对时序依赖性进行建模，甚至可以捕捉到人类关节的**共同现象**。

论文69提出了一种具有dropout的端到端的深度LSTM网络模型；论文18提出了P-LSTM模型以学习每个组别内的身体部位轨迹的长期依赖模式；论文36提出了具有信任门的时空LSTM架构。

### C 居于骨架和depth的结合进行行为识别

两种特征的结合能够克服**人与物之间存在交互**或者**行为间具有非常相似的运动轨迹**。

论文12提出，通过结合**四种类型的局部特征（来自depth和节点坐标）**来处理局部遮挡问题同时增加识别准确度。

论文47提出，其中局部占用模式Local Occupancy Patterns，HON4D和基于骨架的特征与分层混合范式相结合，这些规范化了的每个身体部位在每个模式中的权重。

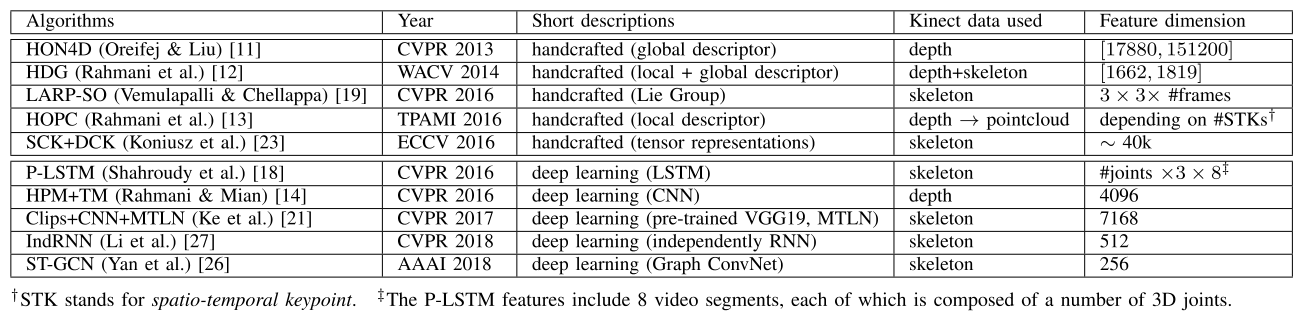
为了提高模型的性能，论文79对生成的depth和骨架特征采用了双线性压缩层技术。

论文72使用典型的相关性分析以最大化从不同的感知器中提出到的特征之间的相关性。

以上论文中的特征都包含从骨架数据中提取得到的角度信息、depth视频中的depth运动图以及RGB视频中的视觉流。

**对所有特征共享的子空间进行学习，以及采用平均池化的方式，得到最终的特征描述子**。

## 3 分析和评估算法



他们可以使用的类型由四种：基于骨架、基于depth、手工制作和深度学习得到的特征。

### 1 HON4D

论文11提出了全局特征描述子以捕获行为的几何关系以及运动信息。值得注意的是这些信息来自4D空间：二维的空间坐标、时间和depth。而为了构建该描述子，作者使用了一个由120个顶点和600个单元的多色球来量化的空间。

**将从远点到这些顶点的向量作为投影轴，以得到每个视频序列的法线分布。**

为了提升分类器的性能，随机的干扰random perturbations施加在了这些投影上，另外，HON4D的特征维度会随着数据集的不同而变化。

### 2 HDG

论文12中，每个depth序列首先被分割成小的子空间subvolumes；其次对每个子空间计算其对应的depth的分布直方图和depth的导数。而对于每个骨架序列，骨架中的躯干关节torso joint被作为一个稳定的参考点，以计算节点坐标差异的分布直方图。另外，将个节点运动量的变化纳入全局特征向量以形成**时空节点特征**。

两个随机决策森林中，其中一个进行特征剪枝，另一个用于分类。关于特征维度和特征剪枝的细节将在第四部分介绍。

### 3 HOPC

论文13将depth图片建模成3D点云的形式。

作者使用两种类型的支持量，即所谓的空间支持量（支持空间）和时空支持量。

HOPC描述子是从每个点周围的支持空间内的点云中提取出来的，如果点云的特征值比大于某个预先定义的阈值，则可以将点云分类为时空关键点STK。

而对于每个STK，算法进一步将特征向量投影到正十二面体的20个顶点的轴上。

而最终的每个STK的HOPC描述子是由三个小的直方图concatenate而成。而每个直方图都捕获了点云在支持空间内的特征向量的分布。

### 4 LARP-SO

论文19通过SO将论文17的工作做了扩展，作者称其为LARP-SO算法，该算法中，the rolling map 用来描述黎曼流形如何沿着平滑的滚动曲线在另一个流形上滚动。

每个骨架序列通过不同身体部位之间的相对3D**旋度**来表示，同时每个行为被表示为一个曲线。由于很难在非欧几里得空间中perform行为的曲线，因此the curves 通过对数映射展开，同时rolling map用来减少变形distortions。

**傅里叶时序金字塔表示在本算法中用来构造描述子以对噪音更具有鲁棒性，同时对时序错位问题不过度敏感**。

### 5 SCK+DCK

论文23使用tensor表示以捕获3D人类节点间的高层关系。算法使用了两个不同的RBF核，参考SCK序列兼容性核和DCK动态兼容性核。前者用来捕获节点的时空兼容性，而后者构建了一个动态行为序列的模型。最后SVM会基于线性特征图进行训练以进行行为识别。

### 6 HPM+TM

论文14应用了一个字典，它包含了捕获到的典型的人类姿势。

一个深度的CNN框架用来训练一个具有视角不变性的姿势模型。与LARP-SO相似的是，傅里叶时序金字塔用来捕获动作视频中的时序特征。而最终的描述子是所有segments的傅里叶系数的集合。

### 7 P-LSTM

论文18将躯干关节的3D坐标从相机坐标体系转换到躯干坐标体系，同时将原点设置在脊椎。其次，将身体其他部位的节点坐标根据hip节点和脊柱节点的距离进行缩放。

然后，**通过将LSTM模型中的记忆细胞分割成基于身体部位的子细胞**，以建立P-LSTM模型。对于每个视频序列，预处理的人类节点被分成5个部分，躯干、两只手和两条腿，同时视频被分成8个同等大小的视频segments部分。

其次对于每个视频片段中的**随机选择出的帧**，其中的**节点坐标通过合并concatenation**，同时被送入模型以学习普遍身体每个部位的时序模式，再将他们结合成一个全局表示。

### 8 Clips+CNN+MTLN

论文21提出一种称为片段的骨骼表示。

首先，将人体关节的笛卡尔坐标转换成圆柱坐标以生成三个clip片段，同时每个clip片段对应圆柱坐标的一个通道。

而为了对整体视频序列的时序信息进行编码，**四个稳定的节点将被作为参考节点**以生成四个坐标系。

预训练好的VGG19网络被用作特征提取器，以学习由该四个coordinate frames中的长期时空特征。

进而，论文21提出应用多任务学习网络MTLN以结合CNN特征中的空域结构信息。

### 9 IndRNN

论文27提出了一种新的RNN模型，和独立的循环神经网络，其中每一层的神经元独立于其他的神经元，但是他们在不同层之间得到fuse融合。

最终对个IndRNNs网络可以通过stack堆叠成一个深度的网络模型。

### 10 ST-GCN

论文26提出了时空图表示，它是GCN的一个延伸，专门处理行为识别的问题。

首先，时空图通过在相邻的节点之间和时序方向的节点之间的增加边而构建成。

因此，GCN和一个分类器将被用来推断图中的依赖性。

## 4 实验设置

六个算法的代码来自开源，而剩下的四种，作者使用matlab自行实现。此外，作者使用了HDG的十种变体以评估“每个描述子的组合方案”的性能。

作者将传统的RNN和LSTM模型视为baseline方法，同时添加了四个P-LSTM的变体，以评估使用不同数量的视频片段对性能的影响。

### A Benchmark 数据集

#### 1 MSRAction3D

20个运动相关的行为。

每个行为被10个人表演2-3次。

难度在于高度的人物交互性。

#### 2 3D Action Pairs

6对具有相似运动轨迹的动作。

每个动作被10个人表演三次。

难度在于1 每对行为中包含相似的运动轨迹；2 与subject互动的物体object仅仅能被呈现在RGB D数据中，而不能展现在骨架中。

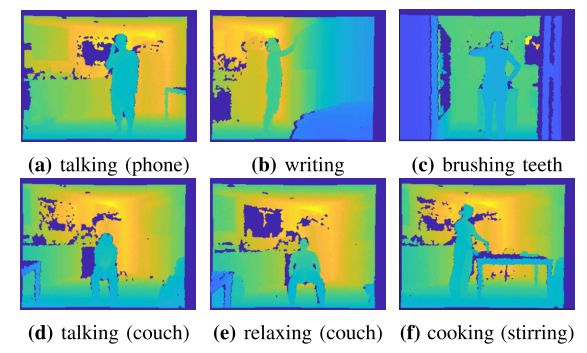
#### 3 Cornell Activity Dataset

包含两个子数据集。

每个数据集，均包含了RGB数据和可追踪的骨架数据。

本次实验中仅仅CAD-60类型被使用。

CAD数据集见下，同时它展现了depth视频中的高等级的噪音。



#### 4 UWA3D Activity Dataset

它包含30个动作，由10个不同身高的人在杂乱的场景中以不同的速度执行

难度在于其高度的类间相似性和频繁的自我遮挡行为。

#### 5 UWA3D Multiview Activity II

30个动作，由9个人在杂乱的场景中执行。

该数据集中Kinect相机被移动到不同的位置以捕获四个不同视角的行为。

因为多视角的性质因而其更具有挑战性。

#### 6 NTU RGB+D Dataset

现今最大的基于Kinect相机的数据集，它包含56880个视频片段和超过400万张的帧。

包含60个行为，由40个人从80个views执行行为。

数据集包含不同的序列长度，同时展示出**高度的类间差异性**。

**使用数据集的方式**

有四个数据集用于多任务单视角的实验。

对于每个数据集，我们使用一般的subject数据用于训练，剩下的用于测试。

我们测试所有可能的subject在训练集和测试集上的组合，以计算每个算法的平均准确率。

对于一个具有10个人物的数据集，有252个实验被执行。

**The UWA3D Multiview Activity II dataset**

它用于多视角的实验，两个视角用于训练，剩下的用于测试。关于视角共有12个不同的视角组合。

**The NTU RGB+D dataset**

它既用于多subject，又用于多视角 的实验。

尽管indications表明有80个视角用于行为识别，但是数据样本根据三个相机的设置被分组。

对于多视角任务使用两个相机的样本用于训练，剩下一个用于测试。

总的来说，相机在训练集和测试集上的不同组合有3种。

### B 评估设置

#### 1 HON4D

每一帧的大小是320\*240；每个视频被分成4\*3\*3（width\*height\*frames）个时空单元。

#### 2 HOPC

论文13使用不同的**时空尺度。而本文中将使用一个固定不变的scale。**

有两个数据集的时序尺度设置为2而空域尺度设置为19。

对于其他数据集，我们使用2作为时序尺度、140作为空域尺度。

进而，我们将depth视频分割成6\*5\*3个时空单元，对应X\Y和time。

#### 3 LARP-SO

本模型中使用的最佳的帧数因数据集的不同而不同。一些数据集的帧数设置为100，一些则是76、111和1000。

#### 4 SCK+DCK

使用了作者的最新模型，它不仅融合序列信息，还有子序列信息。

SCK，作者标准化了**所有节点相对于整个框架的髋关节以及所有身体部位的长度。**

DCK，作者未归一化节点，同时假定节点坐标的偏移，能够捕捉到他们的时间演化。

#### 5 HDG

子空间的数量对辨识力的特征无明显影响；

作者将视频片段分割成10\*10\*5个子subvolums，（X\Y\time）以计算depth数据的分布直方图。

对于节点移动volume特征，每个节点的volume被分割成1\*1\*5个单元。

有四个独立的特征表示简述encapsulated为：

1 depth分布直方图、2 depth梯度的直方图、3节点位置差异differences、4 节点移动volume特征。

#### 6 HPM+TM

将傅里叶金字塔的层级设置为3，而低频率的傅里叶系数设置为4。

作者使用论文14提出的模型以获取到视角不变特征说，我们**比较了基于平均池化和时序模型的CNN特征提取算法的**识别精度。

#### 7 P-LSTM

作者使用与论文18相似的规范化处理步骤。

视频部分的数量和模型各层中的隐藏单元的数量被设置为8或者50。

实验中学习率设置为0.01，max\_epoch等于300.

#### 8 Clips+CNN+MTLN

学习率设置为0.001，MTLN中的batch\_size设置为100。

同时我们选择了四个而不同的实验设置以对比其识别的性能。

#### 9 IndRNN

我们使用了学习率为0.0002的Adam优化器，以及如果评估准确率如果没有增加将会使学习率降低10倍。

对于subject和view的实验，dropout分别设置为0.25和0.1.

#### 10 ST-GCN

对于卷积操作，我们根据ablation study烧蚀实验，使用了最优的分组策略partition strategy。由于不同的数据集有不同的节点，我们因而也构建了时空骨骼图。

**对于剩下的五个数据集，我们使用了不同的设置，因为每个视频只有一个表演对象（人物）。**

### C 评估指标

全局平均准确率；各类别识别准确率（混淆矩阵）；

展示了每个算法在M个数据集上的全局性能，且表示第个数据集上的排序值。而对于六个算法的比较是通过平均rank来表达。

### D HDG超参数的优化

暂不考虑，太远了!

## 5 实验结果