## VNect:通过单一的RGB相机进行实时的3D人体姿态估计

### 概要

本文提出了第一种使用单一RGB相机以稳定，时间一致的方式捕获人体全部3D骨骼姿态的实时方法。本文的方法结合了一种新型的卷积神经网络(CNN)，这种神经网络基于采用了运动骨骼拟合的姿态回归。本文中新型的全卷积姿态公式实时地结合回归2D和3D关节位置，因此不需要紧凑的裁剪输入框架。实时运动骨架拟合方法使用CNN输出在相干运动骨架的基础上产生时间上稳定的3D全局姿势重建。这使得本文的方法-首次单一RGB相机方法，可以运用于实时的应用中，例如：3D人物控制-至今为止，对于此应用的唯一单目相机的方法就只有专业的RGB-D相机。本文的方法的精确度在数量上与最佳离线3D单一RGB姿态估计方法相当。该方法的结果在质量上与单一RGB-D方法（例如微软的Kinect）的结果相当，有时甚至更好。然而，本文发现该方法比RGB-D方案还有更广阔的通用性，例如：户外场景的工作，社区录像，以及低质量的日用RGB相机。

**关键词**：人体姿态，单目，实时

1介绍

人体的光照骨骼运动捕捉被广泛的应用于一些应用中(例如电影和游戏的角色动画，体育和生物力学，以及医学)。为了去克服需要标记套装的商业系统所强加的可用性约束，研究者开发出了更少标记点的运动捕捉方法，这些方法使用多视图视频在更一般的场景估计姿态,最近的解决方案是实时的。应用程序(例如实时运动驱动3D游戏人物控制，自我沉浸式3D虚拟和加强现实以及人机交互)的普及，已经产生了新型的实时全身体的运动评估的技术，这种技术仅仅使用了一个简单的，容易安装的，具有深度的相机（例如：微软的Kinect）。RGB-D相机提供了有价值的深度信息，这些深度信息可以极大的简化单目姿态重建。然而，RGB-D相机经常在一般的室外场景失败（由于光线影响）是非常糟糕的，而且还有更高的能量消耗，更低的分辨率以及受限制的范围，甚至不能像彩色相机那样广泛和便宜。大多数方法也只是捕获相对于一个边界框的局部3D姿态，并不是全局的3D姿态。这使得他们不适用于一些应用（例如实时的3D人物控制)。

在这篇论文中，他们提出了第一种捕获时间一致的全局3D人体姿势的方法 -在一般环境中从单个RGB视频实时（30 Hz）实现单个稳定运动骨架的关节角度。这种方法使用了卷积神经网络。为了去提高实时性，他们使用了一个更浅的50层网络。然而，为了在实时框架上的最好的质量效率，他们不仅仅使用了一个更浅的变种，而是把它扩展为一个新型的全卷积神经公式。这使得更高精确度的2D和3D的姿态回归，特别是末关节（例如：手，脚），在实时上。与现有解决方案相比，他们的方法允许对非裁剪图像进行操作，并且在运行时受到关注的情况下，它可以用于引导简单的边界框跟踪器。他们还将基于CNN的关节位置回归与有效的优化步骤结合起来，以时间稳定的方式将3D骨骼拟合到这些重建，从而产生骨骼的全局姿势和关节角度。总之，他们通过提出第一个从单个RGB视频捕获全局3D运动骨架姿势的实时方法做出贡献。

**2 相关工作**

他们的目标是通过一个单一相机实时的捕获稳定的3D骨架运动。他们重点讨论了大量无标记运动捕获研究方法的相关工作，这些研究有助于实现这些特性中的任何一个。

多视图：通过多视图设置，无标记运动捕捉解决方案可实现高精度。大多数方法都通过离线计算来实现高质量。通过提出高斯以及其他的近似方法来获得实时的表现。然而，这些基于跟踪的方法经常在它们优化的非凸拟合函数的局部最小值中失去跟踪并且需要单独的初始化。通过生成和判别估计的组合可以提高稳健性，甚至可以从单一输入视图和自我中心视角。他们利用生成跟踪组件来确保时间稳定性，但通过完整的图像形成模型避免模型投射以加速估计。相反，他们将判别式姿势估计与运动学拟合相结合，以在欠约束环境中取得成功。

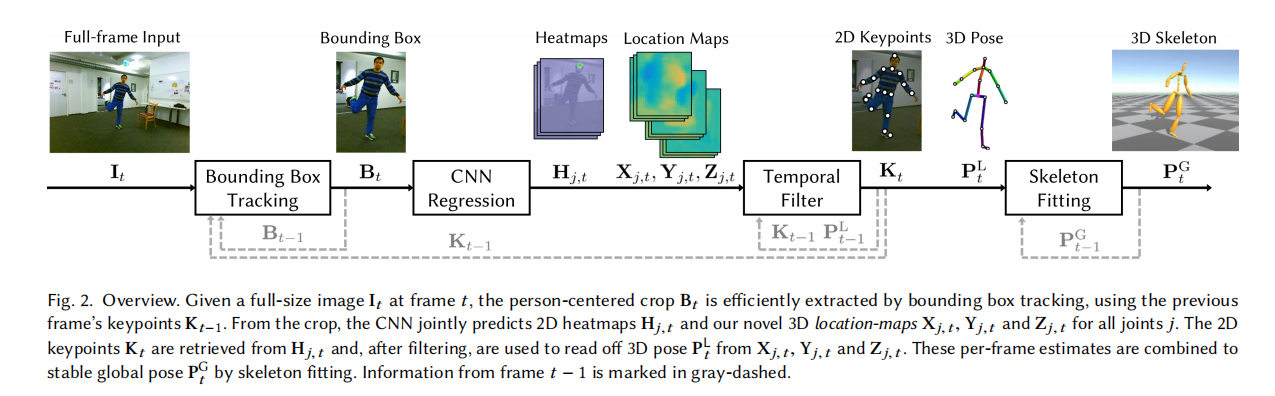
基于单眼深度：RGB-D传感器提供的额外深度通道带来了强大的实时姿态估计解决方案而且低成本设备的可用性已经实现了一系列新应用。甚至已经证明了对一般变形物体的实时跟踪和无模板重建。RGB-D信息克服了单目姿势估计中的前向模糊。本文的目标是视频解决方案，这个方案在不依赖专用有源传感器的情况下克服深度模糊。

单目RGB：单眼生成动作捕捉仅用于短片，并与强动作前置配对或与判别性重新初始化相结合，因为生成性重建从根本上是不受约束的。使用照片般逼真的模板模型进行模型拟合可以实现更简单的单眼跟踪简单运动，但需要更昂贵的离线计算。基于抽样的方法避免了局部最小值。然而，由于一个受限制的样本数量，一个实时变种不能确保全局收敛，例如：粒子群技术。运动结构技术利用一批帧中的运动提示，并且还应用于人体运动估计。然而，批优化并不适用于本文中实时的帧顺序流传输环境。然而对于一些应用，手动注释和帧的校正是合适的，例如使电影演员在视频中重塑和更换。由于物理的约束，通过单目录像的高精确度地重构是可能的。Vondark等人通过模拟Biped控制器而无需手动注释即可成功，但需要批量优化。虽然这些方法可以产生高质量的重建，但是交互和昂贵的优化排除了实时应用。

类似地，本文中提出了一种3D姿态估计方法，该方法利用新颖的完全卷积CNN公式来联合预测2D和3D姿势。结合廉价的预处理和基于优化的骨架拟合方法，它可以实现高精度的姿态估计，同时以超过30 Hz的频率运行。

**3 概述**

本文的系统能够从单目RGB相机获得人类在时间上一致的全3D骨骼姿势。从单个RGB相机估计3D姿势是具有固有模糊性的具有挑战性，欠约束的问题。图2概述了本文解决这一具有挑战性问题的方法。它由两个主要组成部分组成。 第一个部分是卷积神经网络（CNN），用于在不适定的单眼捕获条件下回归2D和3D关节位置。它在已知的3D人体姿势数据集上进行训练，另外利用带注释的2D人体姿势数据集来改善户外性能。第二个部分将回归的关节位置与运动骨架拟合方法相结合，以产生时间稳定的，相机相对的全3D骨骼姿势。



CNN姿态回归：本文的方法的核心是CNN，它实时预测2D和根（骨盆）相对3D关节位置。新提出的完全卷积姿势公式导致结果与3D关节位置精度的最新离线方法相同。完全卷积，它可以在没有围绕主题的紧凑裁剪的情况下操作。无论场景设置如何，CNN都能够预测各种活动的关节位置，为进一步的姿势细化提供强大的基础，以产生时间上一致的全3D姿势参数。

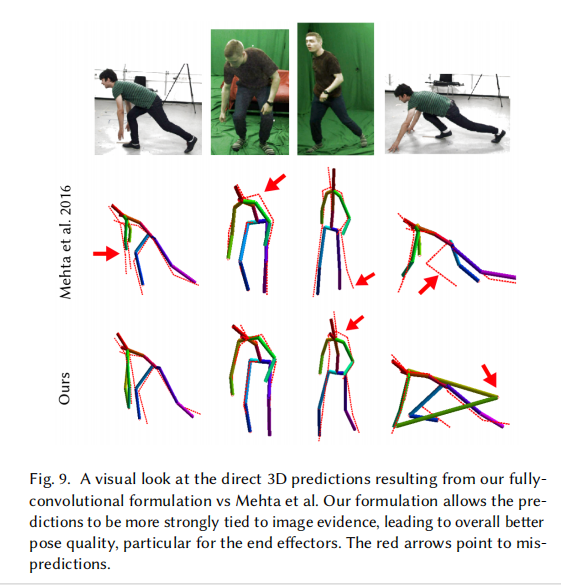
运动骨骼拟合：可以利用来自CNN的2D和3D预测以及序列的时间历史来获得时间上一致的全3D骨架姿势，其中骨架根（骨盆）定位在相机空间中。本文的方法使用以下优化函数：（1）将预测的2D和3D关节位置组合以适合最小二乘意义上的运动骨架，（2）确保随时间的时间上的时间平滑跟踪。通过在不同阶段应用滤波步骤来进一步提高跟踪姿势的稳定性。

**4 实时单目RGB姿态估计**

## **4.1 CNN姿态回归**

CNN姿态回归的目标是去获的2D以及3D空间的关节位置。对于使用了神经网的2D姿态回归，从x，y体关节坐标到基于热图的体关节检测公式的直接回归的公式的变化一直是2D姿态估计的最新发展背后的关键驱动因素。基于热图的公式通过预测每个关节j∈{1..J}的图像平面上的置信热图Hj,t，自然地将图像证据与姿势估计联系起来。

现有的3D姿势估计方法缺乏这样的图像到预测关联，通常直接回归根相对关节位置，导致预测姿势的额外关节不能反映图像中人的关节。如图9所示。将姿势作为关节位置的矢量处理也导致对具有完全连接配方的网络的自然引力，限制以固定分辨率对紧密作物的输入，这是需要克服的限制。这些方法假定紧密边界框的可用性，这需要补充用于实际使用的单独的边界框估计器，这进一步增加了这些方法的运行时间。Pavlakos等人的完全卷积式公式试图减轻其中的一些问题，但受到昂贵的每关节体积公式的限制，这种公式仍然依赖于裁剪的输入，并且不能很好地扩展到更大的图像尺寸。

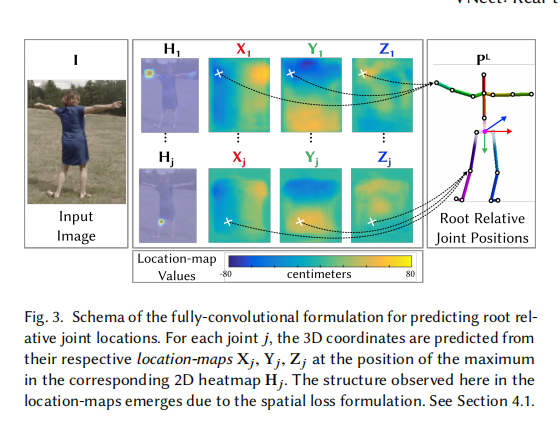


本文通过新的公式克服了这些限制，通过使用每个关节j的三个附加位置图Xj，Yj，Zj将2D热图制剂扩展到3D，分别捕获根相对位置xj，yj和zj。为了使3D姿势预测更强地链接到图像中的2D外观，xj，yj和zj值从它们各自的位置图中读出，位于相应关节的2D热图Hj的最大值的位置处，并存储在 PL = {x，y，z}，其中x∈R1\*J是存储每个关节最大值的坐标x位置的向量。这个姿态公式如图3 所示。使用这个全卷积公式的神经网络对输入图像的大小没有约束，不需要紧凑的剪裁也可以工作。此外，网络提供2D和3D联合位置估计而没有额外的开销，该方法在后续步骤中利用这些开销进行实时估计。

损失期限：为了强制该方法只对来自联合j的2D位置的各自地图的xj，yj和zj感兴趣，联合位置图损失在关节的2D位置周围加权更强。使用L2损失。 对于xj，损失公式是：

 (1)

GT表示基本事实，⊙是Hadamard产品。位置图用相应的地面实况2D热图HjGT加权，其依次具有等于高斯的置信度，其中小支撑位于关节j的2D位置。请注意，位置图上没有结构。在预测的位置图中出现的结构表示xj和yj与图像平面中的关节j的根相对位置的相关性。见图3。

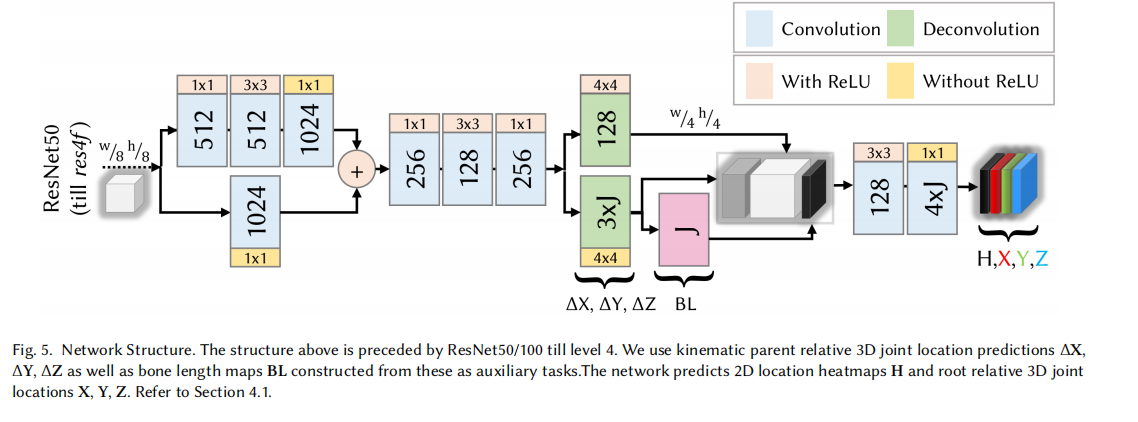


网络详细信息：使用提议的公式来调整He等人的ResNet50网络架构。将res5a的ResNet50的层数开始替换为图5所示的架构，生成所有关节的热图和位置图j∈{1..J}。训练之后，批量标准化层与其先前卷积层的权重合并，以提高正向通过的速度。

中级监督：根据res4d和res5a的特征预测2D热图和3D位置图，随着迭代次数的增加逐渐减少中间损失的权重。另外，类似于根相对位置图Xj，Yj和Zj，预测来自res5b处的特征的运动学父亲相对位置图ΔXj，ΔYj和ΔZj，并计算骨骼长度图：

 （2）

随后将这些中间预测与中间特征连接起来，以给网络一个明确的骨长度概念来指导预测。见图5。



## **4.2运动骨架拟合**

在视频上应用每帧姿态估计技术不利用并确保运动的时间一致性，并且小的姿势不准确导致时间抖动，这对于大多数图形应用来说是不可接受的伪像。该方法将联合优化框架中的2D和3D关节位置与时间滤波和平滑相结合，以获得准确，时间稳定且稳健的结果。首先，2D预测Kt在时间上被过滤并用于从位置图预测中获得每个关节的3D坐标，给出PtL。为了确保骨骼稳定性，在一个简单的重新定位步骤中，PtL固有的骨骼长度被底层骨骼的骨骼长度所取代，这一步骤保留了PtL的骨骼方向。通过最小化目标能量来组合所得的2D和3D预测:

 (3)

用于骨骼关节角度θ和根关节在相机空间中的位置d。3D逆运动学术语EIK通过与3D CNN输出PtL的相似性来确定整体姿势。投影项Eproj确定全局位置d并通过重新投影到检测到的2D键点Kt来校正3D姿势。这两个公式都是通过L2损失来实现的，

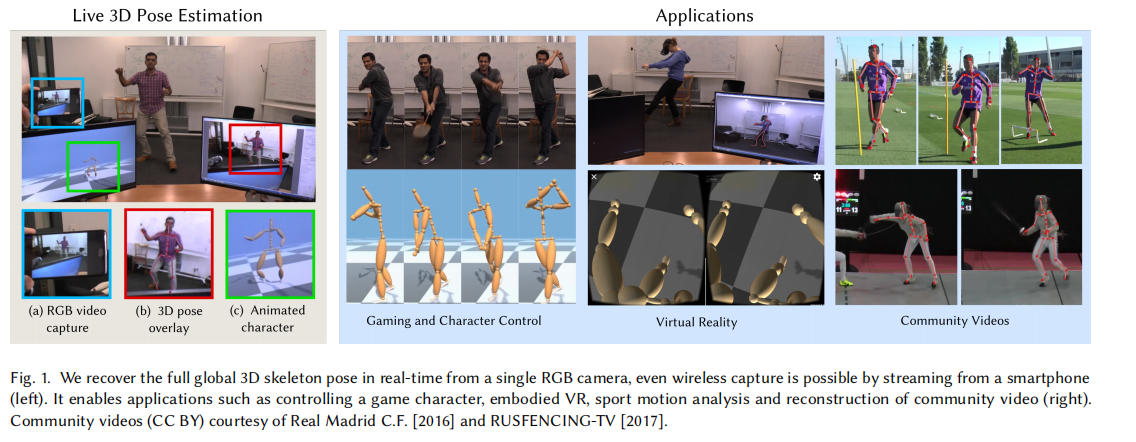
 （4）

其中Π是从3D到图像平面的投影函数，PtG = PtG（θ,d）。 假设针孔投影模型。如果摄像机校准未知，则假定垂直视场为54度。在Esmooth=∥PtG∥2之前，使用平滑度来强制执行时间稳定性，从而惩罚加速度PtG。 为了抵消单眼重建中的强烈深度不确定性，该方法还使用Edepth =∥[PtG]Z∥2来惩罚深度的大变化，其中[PtG]Z是3D速度PtG的z分量。最后，还使用1Euro过滤器过滤3D姿势。

参数：能量项EIK，Eproj，Esmooth和Edepth分别用ωIK= 1，ωproj= 44，ωsmooth= 0.07和ωdepth= 0.11加权。1Euro滤波器的参数根据经验设置为fcmin = 1.7，β= 0.3用于滤波Kt，fcmi = 0.8，β= 0.4用于PtL，并且fcmi= 20，β= 0.4用于滤波PtG。我们的实现使用Ceres库中的Levenberg-Marquardt算法。

**5结果**

该方法以30 Hz的频率显示系统的实时应用。重建质量很高，本文展示了他们的方法3D角色控制，体现虚拟现实和低质量智能手机相机流的姿势跟踪的实用性。见图1。在补充视频中最好观察运动结果。使用视频解决方案实现这些应用的步骤的重要性在10多个序列中得到了彻底的评估。结果在质量上与基于深度相机的解决方案（如Kinect）相当并且明显优于现有的单眼视频解决方案。作为定性基线，他们选择了周等人最先进的2D到3D提升方法。以及Mehta等人的3D回归方法，它可以离线估计关节位置。在已建立的H3.6M数据集和MPI-INF-3DHP数据集上进一步定量验证了准确度的提高。几个实时示例和社区视频演示了对不同人物，服装和场景的稳健性。



**6结论**

他们已经提出了第一种方法，该方法以30Hz的单个RGB视频流以稳定的，时间上一致的方式估计人的3D运动姿态，包括全局位置。他们的方法结合了回归2D和3D关节位置的完全卷积CNN和运动骨架拟合方法，产生运动的实时时间稳定3D重建。与大多数现有方法相比，本文的网络可以在没有严格边界框的情况下运行，并且便于廉价的边界框跟踪。他们已经在各种具有挑战性的实时场景中展示了结果，包括智能手机摄像头的实时流媒体以及社区视频。已经提出了许多应用，例如用于计算机游戏的体现VR和交互式角色控制。

定性和定量评估表明，本文的方法与离线的最先进的单目RGB方法相比，并且接近实时RGB-D方法的质量。因此，相信他们的方法是向3D人体姿态估计进行民主化的重要一步，同时满足了对基于IR的深度相机等特殊相机的需求，以及长而重的处理时间。