发布时间：2017
场景 :实时（30Hz)，视频，
计算资源 : Titan X
普遍效果 : MPJPE, Res100 82.5,res50 80.5 on human 3.6, res50 MPJPE 125.7 on MPI-INF-3DHP
核心要点 : fully CNN+skeleton fitting

Abstract

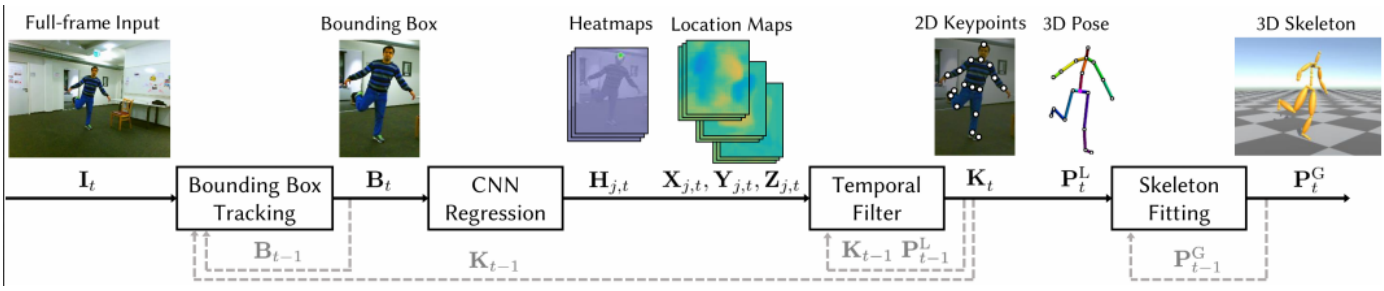
fully-convolutional pose formulation regresses 2D and 3D joint postitions jointly in realy time and does not require tightly copped input frames
the first monocular RGB method usable in real-time applications(没有用深度信息）, the accuracy is on par with the best offline 3D mononcular RGB PE.

1 Introduction

Based on the RGB-D cameras, skeletal pose estimation from a single color camera is challenging.
在当时（2017），基本上都是在做2D, 3D都是线下一张图片一张图片的处理，或者在一个界定好的框里，这些都不适用于实时。
The article solution combines:

- A new real-time, fully-convolutional(shallower but effective) 3D body pose formulation using CNNs that yields 2D and 3D joint postitions simultaneously and forgoes the need to preform expensive bounding box computations.(因为找bounding box耗费时间)
- Model-based kinmatic skeleton fitting against the 2D/3D pose predictions to produce **temporally stable joint angles of a metric global 3D skeleton** , in real time.

2 OverView



As the Figure2 shows, out method consists of two primary components:

- **a convolutional neural network to regress 2D and 3D joint positions under the ill-posed monocular capture conditions**: Because the fully-convolutional, it can operate in the absence of tight crops around the subject. and is capable of predicting joint positions for a diverse class of activities regardless of the scene settings.
- **combines the regressed joint positions with a kinematic skeleton fitting method to produce a temporally stable , camera-relative,full 3D skeletal pose**:

    1. combines the predicted 2d and 3D joint positions to fit a kinematic skeleton in a least squares sense
    2. ensures temporally smooth tracking over time
- Skeleton Initialization

3 Real-Time Monocular 3D pose Estimation

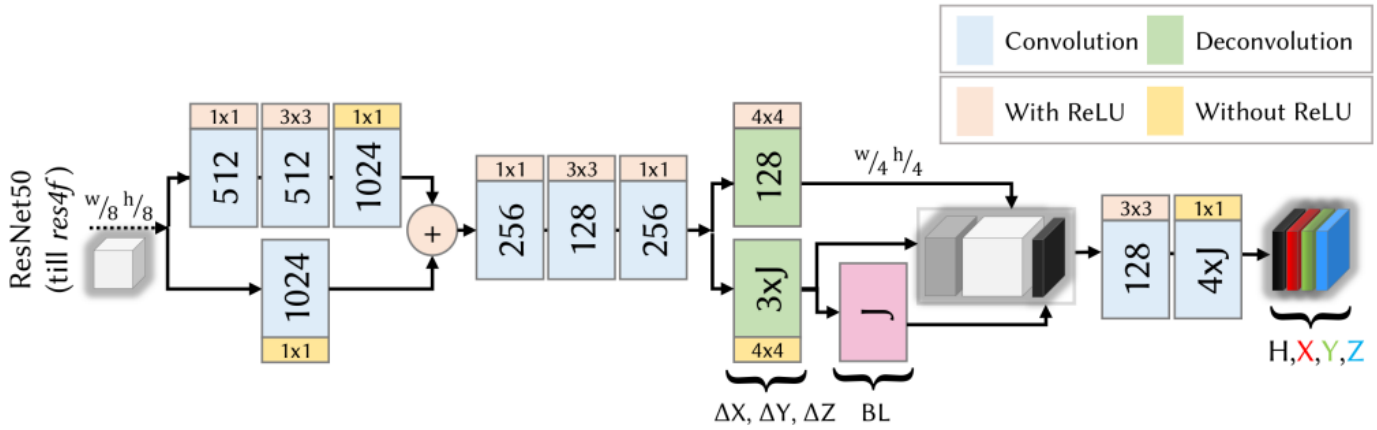
As inputs, a coninuous stream of monocular RGB images $\{\ldots, I_{t-1}, I_t\}$, the 2D jonit positions estimate K_t, the root-relative 3D jonit positions P_t^L and the final output is $P_t^G(\theta, d)$ which is prameterized by the global position d in camera space and joint angles θ of the kinematic skeleton S.

3.1 CNN Pose Regression

现有的3D预测的方式缺少直接图片到预测的联系，通常是对root-relative joint locations回归，导致预测姿势不正确，且需要bounding box。
我们在直接预测的基础上，通过对每个关节使用三个额外的location-maps X_j, Y_j, Z_j将2D热图扩展到3D(也可以视为对每个关节实现root_relative joint locatoin). 其中X表示X坐标的最大值，Y存储Y坐标的最大值。
the Loss: for x_j, the loss formulations is:

$$Losss(x_j) = \|H_j^{GT} \odot (X_j - X_j^{GT})\|_2$$

其GT表示ground truth, H^{GT}表示ground truth 2D heatmap
Network Details: adapt the resNet50, replace the layers of ResNet50 from res5a onwards with the architecture depicted in Figure5:



Intermediate Supervision：中间监督总共有三个部分，最后结合：

- 2D feature map
- computed the bone length maps(见上图中间）：
  $$BL_j = \sqrt{\Delta X_j \odot \Delta X_j + \Delta Y_j \odot \Delta Y_j + \Delta Z_j \odot \Delta Z_j}$$
- 对bone length maps做Batch normalization **Bounding box tracking**: 虽然bounding box的寻找耗费时间，但是使用bounding box也有助于减少CNN的特征提取空间，减少时间，折中，我们在2D预测时顺便预测下一个frame的范围，即比当前预测的范围稍微大一些，这样也不会额外增加很多训练时间。

3.2 Kinematic Skelelton Fitting

单独的处理视频中的每个图片，会造成动作上的不连续，即**抖动, so we combine the 2D and 3D joint positions in a joint optimization framework, along with temporal filtering an d smoothing**.
2D predictions K_t are temporally filtered and used to obtain the 3D coordinates of each joint from the **location-map predictions**, giving us P_t^L.(为了保证骨架的稳定，骨架的长度被之前步骤中定义好的骨架长度替换，只保留方向）， the 2D and 3D predictions are combined by minimizing the objective energy:

$$E_{total}(\theta, d) = E_{IK}(\theta, d) + E_{proj}(\theta, d) + E_{smooth}(\theta, d) + E_{depth}(\theta, d)$$

其中:
θ: skeletal joint angles
d: root joints location in camera space
E_{IK}: the 3D inverse kinematics term determines the overall pose by similarity to the 3D CNN output P_t^L, implemented with the L2 loss: $E_{IK} = \|(P_t^G - d) - P_t^L\|_2$

E_{proj}: the projection term determines global position d and corrects the 3D pose by re-projection onto the detected 2D keypoints K_t. implemented with the L2 loss: $E_{proj} = \|\prod(P_t^G) - K_t\|_2$,其中$\prod$是从3D到图片平面的映射函数。

$E_{smooth} = \|\widehat{P_t^G}\|_2$: penalizaing the acceleration $\widehat{P_t^G}$, assumed as the smootheness prior.

$E_{depth} = \|[\widetilde{P_t^G}]_z\|_2$:the z component of 3D velocity $\widetilde{P_t^G}$ counteract the strong depth uncertainty in monocular reconstruction, penalize large variations in depth.

4 Results

4.1 Comparison with Active Depth Sensors(Kinect,RGB-D)

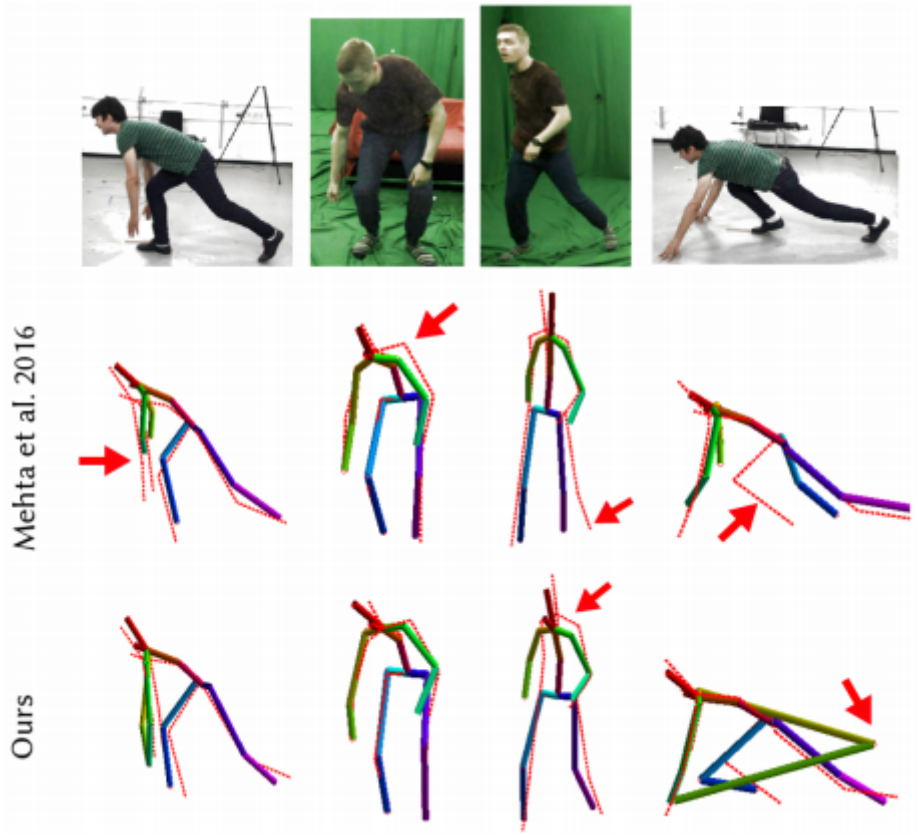
如下图，分别是室内和室外的对比，论文没有深度信息的实现比kinect还好一些,在室内四肢能和背景区分开。

Fig. 6. Side-by-side pose comparison with our method (top) and Kinect (bottom). Overall estimated poses are of similar quality (first two frames). Both the Kinect (third and fourth frames) and our approach (fourth and fifth frames) occasionally predict erroneous poses.

Fig. 7. Our approach succeeds in strong illumination and sunlight (center right and right), while the IR-based depth estimates of the Microsoft Kinect are erroneous (left) and depth-based tracking fails (center left).

4.2 Comparison with Video Solutions

与当时最好的线下模型Mehta对比(不能实时），论文的模型是30HZ，可以达到实时。从下图中可以发现论文的模型能捕捉到更多的关节细节。


Mehta et al. 2016
Ours