# 人体姿态估计综述

# Part1：Single Person Pose Estimation

2015 年之前的方法都是回归出精确的关节点坐标( x,y )，采用这种方法不好的原因是人体运动灵活，模型可扩展性较差。

#### 《Flowing ConvNets for Human Pose Estimation in Videos》ICCV 2015

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78320781>

2015 年 flow convnet 将姿态估计看作是检测问题,输出是 heatmap。其创新点在于从卷积神经网络的 3 和 7 层提取出来,再经过卷积操作,称之为空间融合模型,用来提取关节点之间的内在联系;同时使用光流信息,用来对准相邻帧的 heatmap 预测。最后使用参数池化方法,将对其的 heatmap 合并成一个 scoremap。

评测数据集：FLIC数据集，对于wrist和elbow的平均PCK可以达到92%，可以做到实时性，速度为5fps。但是该方法对于pose的估计范围有限，只是半身的关节点，并不是全身的身体骨骼点。

#### 《Convolutional Pose Machines》CVPR 2016

延伸阅读:<http://blog.csdn.net/qq_36165459/article/details/78321054>

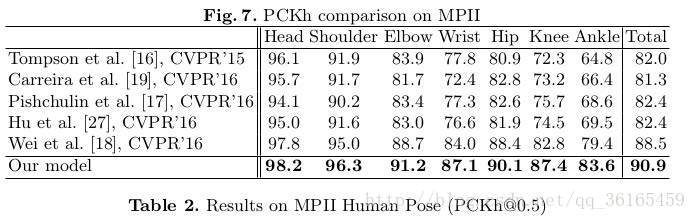
2016 年提出的 CPM 方法具有很强的鲁棒性,之后的很多方法是基于此改进的。CPM 的贡献在于使用顺序化的卷积架构来表达空间信息和纹理信息。网络分为多个阶段,每一个阶段都有监督训练的部分。前面的阶段使用原始图片作为输入,后面阶段使用之前阶段的特征图作为输入,主要是为了融合空间信息,纹理信息和中心约束。另外,对同一个卷积架构同时使用多个尺度处理输入的特征和响应,既能保证精度,又考虑了各部件之间的远近距离关系。

评测数据集：MPII，LSP，FLIC，在MPII数据集上的total PCKh是87.95%（如果加上LSP数据集作为训练，将达到88.52%），在LSP数据集上的PCKh是84.32%（如果加上MPII数据集作为训练，将达到90.5%），在FLIC数据集上的PCK@0.2分别是elbows（97.59%），wrist（95.03%）。速度不明，应该无法做到实时。

#### 《Stacked Hourglass Networks for Human Pose Estimation》ECCV 2016

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78321529>

同年发表的 stacked hourglass 也取得了非常不错的效果。对给定的单张 RGB 图像,输出人体关键点的精确像素位置,使用多尺度特征,捕捉人体各关节点的空间位置信息。网络结构形似沙漏状,重复使用 top-down 到 bottom-up 来推断人体的关节点位置。每一个 top-down到 bottom-up 的结构都是一个 hourglass 模块。

评测数据集：在FLIC数据集上的PCK@0.2分别elbows（99%），elbows（97%）；   
在MPII数据集上如下：Tompson就是flow convnet，Wei就是CPM   


#### 《Structured Feature Learning for Pose Estimation》CVPR 2016

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78321621>

2017 年王晓刚组的 structured pose 也是在 CNN 的基础上进行微调,其创新点在于在卷积层使用几何变换核,能够对关节点之间的依赖关系进行建模,此外还提出了双向树模型,这样每个关节的 feature channel 都可以接收其他关节的信息,称之为信息传递,这种树状结构还能针对多人进行姿态估计。但是这种多人姿态估计的准确度不高，方法还是基于单人的比较好。

评测数据集：FCIL，LSP，MPII，在 FCIL，LSP均比之前的方法有所提升，在MPII数据集上也曾暂列榜首，PCKh达到91.5%，准确率提升不大。

目前在MPII数据集上位列榜首方法来自论文《Adversarial PoseNet: A StructureawareConvolutional Network for Human Pose Estimation》，采用的GAN的方法，效果比之前的state-of-the-art仅仅提升了零点几个百分点。基本上到hourglass之后的方法都是一些微调，虽然理论都不太一样，但是准确度提升不大。

# Part2：Multi-Person Pose Estimation

多人姿态估计的两种研究方法——自顶向下(top-down)：先检测出多个人,再对每一个人进行姿态估计，可以将detection的方法加上单人姿态估计来实现；自底向上(bottom-up)：先检测出关节点,再判断每一个关节点属于哪一个人。

#### 《DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation》 CVPR 2016

#### 《DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model》 ECCV 2016

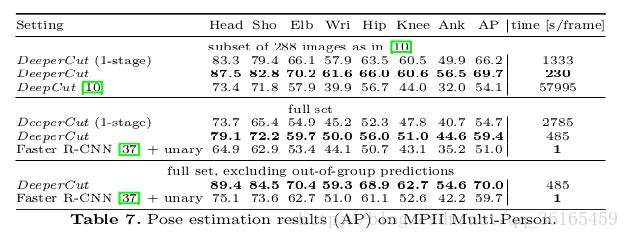
延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78321696>

2016 年的 deepcut,采用自顶向下的方法,先用 CNN 找出所有候选的关节点,将这些关节点组成一幅图,对图中的节点进行聚类,从而判断各个节点属于哪一个人,这是一个优化问题;同时,对各个点进行标记,分类属于身体的哪一部分;两者结合输出姿态估计结果。

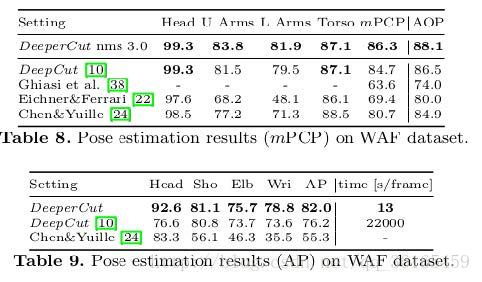
Deepercut 是在 deepcut 的基础上使用 resnet 进行检测提高精度,使用 image conditioned pairwise ,能够将丰富的候选节点进行压缩,提升速度和鲁棒性。

评测数据集：deepcut，对于单人姿态估计，在LSP数据集上的PCK达到87.1%，在MPII数据集上的PCK达到82.4%（可见，适用于多人的姿态估计方法和纯粹的单人姿态估计方法的准确率还有所差距）；对于多人姿态估计，在WAF数据集上mean PCP达到84.7%，在MPII多人数据集上AP 达到 60.5%，速度非常慢。

DeeperCut：和deepcut的评测数据集相同，这里主要针对多人来看，其准确率和速度都有所提升，尤其是速度方面。

在MPII数据集上结果如下：   


可以看到，DeeperCut最快可以做到230s每帧，比deepcut的每帧需要几十万秒速度有显著提升。单纯对于检测来说，faster r-cnn的方法要快很多，不过它的准确度没有deepercut高。

在WAF数据集上也有显著速度提升：   


#### 《ArtTrack: Articulated Multi-person Tracking in the Wild》CVPR 2017

2017年的ArtTrack的作者也是DeeperCut 的第一作者，是将人物姿态估计用到了视频跟踪里面,本文的贡献是利用现有的单帧姿态估计模型作为基础框架,但是速度却明显加快,这种加快主要通过以下两种方式来进行:(1)通过简化和稀疏身体部位的关系图,使用进来的方法进行快速的推理;(2)不加载用于前馈神经网络上的大规模计算量,这些神经网络是为了检测和关联同一人的身体关节。模型仍然是采用 top-down 的方法,即先用 Resnet 检测出body part proposal,然后再根据关联和空间信息将他们归为不同的人。

同时,本文也提出一种 top-down/bottom-up 的模型,即 top-down 部分是用来对人体做一个粗略的估计,之后再用bottom-up 进行精确调整,使得预测的关节点位置更准确。

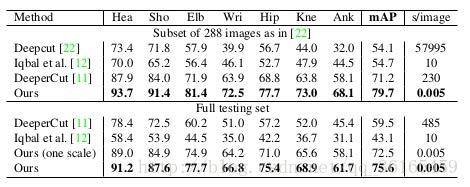
评测数据集：WAF数据集和MPII Video Pose数据集，相应有所提升。

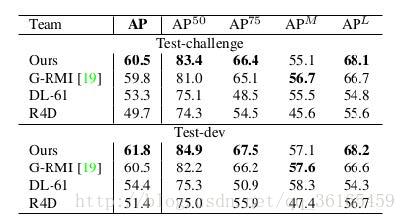
基于Deep(er)Cut和ArtTrack的pose开源实现：<https://github.com/eldar/pose-tensorflow>（python3+tensorflow，MPII数据   
集/COCO数据集）

#### 《Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields》CVPR 2017

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78322184>

2017 年的 Part Affinity Fields(PAF)能够针对多人做到实时检测,它采用的却是自底向上的方法,网络框架分为两路;一路使用 CNN,根据置信图进行关节点预测,另一路使用CNN 获得每个关节点的 PAF,PAF 可以看作是记录 limb 位置和方向的 2D 向量。两路进行联合学习和预测。最后就是如何将这些节点两两连接不重复,这转换为图论问题。

评测数据集：COCO 2016关键点检测数据集+MPII multi-person benchmark。对于MPII多人pose,本文无论是准确度还是精度上都有质的飞跃，其相比于DeeperCut的速度快了4万多倍，准确度也有几个百分点的提升。可以做到实时，每帧只需要5毫秒，即200FPS。   


在COCO2016 challenge上准确度也有显著提升。   


## New state-of-the-Art

#### 《Mask R-CNN》ICCV 2017，FAIR，Kaiming He

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78322340>

2017年何凯明的Mask R-CNN，Mask R-CNN 是用于目标检测分割的框架,即对一张图片,既输出图片中已有的目标,还能为每一个实例生成一个高质量的分割掩码。mask RCNN是在 faster R-CNN 的基础上,在每一个 RoI 都增加一个预测分割的mask,这和分类以及 bounding box 回归是并行的一条分支。它的训练简单,仅仅比 faster RCNN多一点计算开销。它易于泛化到多个任务上,例如人体姿态估计。在不加任何的 trick的情况下,在COCO 数据集上超越其他的方法。因此准确度方面基本上已经是state-of-the-Art。

应用到pose estimation，将分割系统中的目标改为K个one-hot，m\*m的二进制mask。准确率比COCO 2016 冠军高0.9个点，速度达到5 FPS。

#### 《Towards accurate multi-person pose estimation in the wild》CVPR 2017 Google

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78322449>

Google的人体姿态估计，多数时候在论文中简写为G-RMI。

论文采用top-down的结构，分为两个阶段:   
第一阶段使用faster rcnn做detection，检测出图片中的多个人，并对bounding box进行image crop；   
第二阶段采用fully convolutional resnet对每一个bonding box中的人物预测dense heatmap和offset;   
最后通过heatmap和offset的融合得到关键点的精确定位。

#### 《Associative Embedding:End-to-End Learning for Joint Detection and Grouping》

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78322549>

论文提出了一种single-stage，end-to-end的关节点检测和分组方法，这不同于以往的multi-stage的关节点检测方法，在MPII和COCO数据集上达到新的state-of-the-art的效果，超越最近的Mask RCNN和Google GMI。从人体姿态估计方法上属于bottom-up的方法，即先检测关节点，再对关节点进行分组。在COCO测试集上mAP达到0.655。

#### 《RMPE: Regional Multi-Person Pose Estimation》ICCV 2017，SJTU，Tencent Youtu

延伸阅读：<http://blog.csdn.net/qq_36165459/article/details/78330800>

这篇论文是上海交大和腾讯优图的论文，被 ICCV 2017接收。它对于多人姿态估计的方法采用传统的自顶向下的方法，即先检测人，再识别人体姿态。检测使用的是SSD-512，识别人体姿态使用的是state-of-the-art的Stacked Hourglass方法。致力于解决对于imperfect proposal，通过调整，使得crop的单人能够被单人姿态估计方法很好的识别，从而克服检测带来的定位误差。