人体姿态识别研究文献综述

陶逸群[1)]

[1)]杭州电子科技大学 理工类实验班，杭州市 310018

摘 要 近年来，人体姿势识别已逐渐成为计算机视觉领域的研究重点，人们对姿势识别的研究也取得丰厚的成果。本文对 2015 年以来姿识别的研究成果进行了较为系统的综述，并对当前人体姿势识别方法存在的问题以及姿势识别未来的发展方向进行了分析。

关键词 人体姿势识别 深度学习 图像处理 人工智能 关键点检测 CNN 神经网络

中图法分类号 TP391.41 DOI 号

Literature Review Of Human Body Posture Recognition Research

Yiqun-Tao[1)]

[1)]HangZhouDianZi University Excellent College, HangZhou 310018

Summary In recent years, human body posture recognition has gradually become the research focus in the field of computer vision, and people's research on posture recognition has also achieved fruitful results. In this paper, a systematic review of the research results of pose recognition since 2015 is carried out, and the problems existing in the current human pose recognition method and the future development direction of pose recognition are analyzed.

Key words Human Posture Recognition Deep learning Image Processing Artificial Intelligence Key Point Detection CNN Network

0 引言

近年来，随着人工智能技术研究的不断深入，计算机视觉领域也取得了长足的发展，对人的姿态、动作识别和行为理解逐渐成为计算机视觉领域研究的重点。

人体姿态识别的应用范围十分广泛，可用于人机交互、影视制作、运动分析、游戏娱乐以及智能监控等各种领域。人们可以利用人体姿态识别定位人体关节点运动轨迹并记录其运动数据，实现 3D 动画模拟人体运动来制作电影电视；也可以通过记录的轨道和数据对运动进行分析；还可以将人体姿态识别将技术嵌入视频服务器中，运用算法，识别、判断监控画面场景中的动态物体的行为。

因此人体姿态识别具有十分重要的研究意义，对于我国实现现代化建设建设也具有一定的推动作用。

在 2015 年以前，人体姿态识别采用的大都是传统算法——回归出精确的关节点坐标(x,y)。但由于人体运动灵活，导致识别结果的精度较低而且模型的可拓展性也较差。随着深度学习领域的蓬勃发展，现在主流人体姿态的识别算法大都采用了卷积神经网络的方法：通过对大量数据集的训练对人体姿态进行识别。

因此，本文主要是 2015 年之后人体姿态识别

发展的综述，介绍在深度学习领域中不同人体姿态识别方法的优点和不足、目前人体姿态识别领域存在的问题以及未来的发展趋势。

1 人体姿态识别简介

1.1 人体姿态识别概念

人体姿态识别主要在于研究描述人体姿态以及预测人体行为，其识别过程是指，在指定图像或视频中，根据人体中关节点位置的变化，识别人体动作的过程。

1.2 人体姿态识别算法

人体姿态识别的算法只要分为两类，一是基于深度图的算法，另一类是直接基于 RGB 图像的算法。

1.2.1 基于深度图的人体姿态识别算法

深度图是指由相机拍摄的图片，其每个像素值代表的是物体到相机 XY 平面的距离。该算法针对深度图下的人体姿态进行识别。但该算法的应用因采集设备的要求而受限，应用范围较小。

1.2.2 基于 RGB 图像的人体姿态识别算法

基于 RGB 图像的人体姿态识别算法直接通过对红、绿、蓝 3 个颜色通道的变化以及它们相互之间叠加得到的颜色进行识别，不会受到其他因素的干扰限制，因此应用范围广泛，也更具有发展前景。即使是在较为复杂、某种固定的场景中，基于 RGB 图像的人体姿态识别算法相较于基于深度图的人体姿态估计算法也能达到很好的识别效果。

1.3 人体姿态识别的实现

人体姿态是被主要分为基于计算机视角的识别和基于运动捕获技术的识别。

1.3.1 基于计算机视角的人体姿态识别

基于计算机视觉的识别主要通过通过各种特征信息来对人体姿态动作进行识别，比如视频图像序列、人体轮廓、多视角等。

这种识别方式可以比较容易获取人体运动的的轨迹、轮廓等信息，但没有办法具体表达人体的运动细节，以及容易存在因遮挡而识别错误等问题。

1.3.2 基于运动捕获技术的人体姿态识别

基于运动捕获技术的人体姿态识别，是通过定位定位人体的关节点、储存关节点运动数据信息来识别人体的运动轨道。

相较于计算机视角的人体姿态识别，基于运动捕获技术的人体姿态识别可以更好地反映人体姿态信息，也可以更好地处理和记录运动的细节，不会因为物体颜色或被遮挡而影响运动轨道的识别。

1.4 人体姿态识别方法的类型

人体姿态识别的方法有两类，一类是单人姿态估计的方法；另一类的多人姿态估计的方法。单人姿态估计的方法在单人识别效果较好，应用与多人姿态识别的效果还是比较差的。同样，在多人姿态估计效果较好的方法应用在单人姿态估计的效果也不是很理想。

本文将分别对单人姿态识别的方法和多人姿态识别的方法进行介绍。

2 单人姿态识别

2.1 单人姿态识别性能评价指标

对单人姿态识别性能的评价指标为：PCK(Percentage of Correct Keypoints)，即关键点正确估计的比例：检测的关键点与其对应的标签之间的归一化距离小于设定阈值的比例。

对单人姿态识别性能评价的数据集为：MPII 单人数据集、LSP 数据集和 FLIC 数据集。通过对比这三个数据集的 PCK 值来评价模型的好坏。

FLIC 中是以躯干直径作为归一化参考，MPII 中是以头部长度作为归一化参考，即 PCKh。

目前 MPII 单人数据集的部分排名如下表所示:

表 1 MPII 单人数据集排名

Method	PCKh	Method	PCKh
Su,arXiv19	93.9	Chu,CVOR17	91.5
Zhang,arXiv19	92.5	Luvizo,arXiv17	91.2
Tang,ECCV18	92.3	Ning,TMM17	91.2
Ke,ECCV18	92.1	Tang,ECCV18	91.2
Yang,ICCV17	92.0	Newell,ECCV16	90.9
Chen,ICCV17	91.9	Bulat,ECCV16	89.7
Chou,arXiv17	91.8	Wei,CVPR16	88.5

2.2 Flowing ConvNets方法

2.2.1 方法概述

该方法由牛津大学的 Tomas Pfisrer 教授和 Andrew Zisserman 教授以及利兹大学的 Jams Charles 教授在 2015 年 ICCV 会议上发表的《Flowing ConvNets for Human Pose Estimation in Videos》中提

出。

该方法主要采用 CNN 网络来进行人体姿态识别，同时加入了时间信息来提高精度。该方法主要主要有四个创新点：

（1） 提出了一个相比于 Alex-Net 更深的 CNN 网络进行人体姿态估计。而且不同于之前的回归坐标，而是将姿态估计看做是检测问题，输出热力图。这样不仅可以提高关节点定位的鲁棒性，并且更利于在训练过程中的可视化观察。

（2） 提出了一种空间融合层，用来学习隐藏式空间模型，提取关节之间的内在联系。

（3） 使用光流信息，用来对准相邻帧的热力图进行预测。

（4） 使用最后的参数池化层，将对齐的热力图合并成一个聚集的置信图。

2.2.2 网络框架

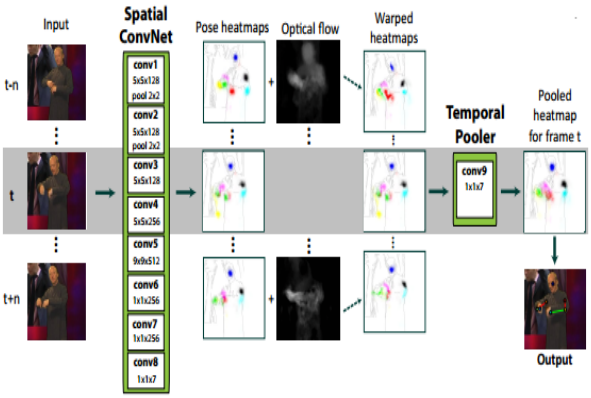


图 1 Flowing ConvNet 网络框架

该方法的网络框架如上图所示。对于当前帧 t，输入它的相邻的前后 n 帧。利用全卷积神经网络（Spatia Net + Spatial Funsion Layers）对每一帧输出一个预测的热力图。再利用光流信息将这些热力图扭曲到当前的帧 t。之后将所有扭曲后的热力图合并到另一个卷积层中，该层学习如何权衡来自附近框架的扭曲的热力图。最后使用集合热图的最大值作为人体的身体关节。

2.2.3 Spatial Funsion Layers 的细节结构

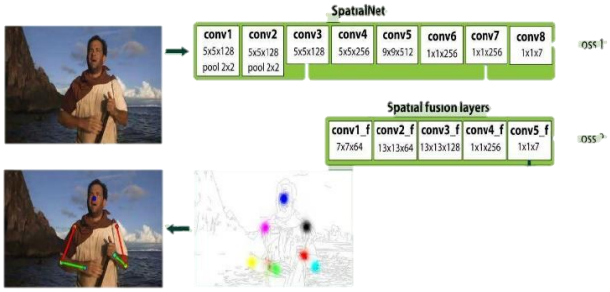


图 2 Spatial Funsion Layers 的细节结构图

Spatial Funsion Layers 的细节结构如上图所示。增加该结构是为了学习关节之间的内在联系。其获取的是之前 CNN 神经网络中的第三个卷积层(conv3)和第 7 个卷积层(conv7)，将这两个卷积层结合之后作为输入，在经过 5 个卷积层。

关于损失函数一共有两个，一个是 spatial net 中的 loss1，采用的是 L2 范式。

loss1 的损失函数公式为：

$$\arg\min_{\lambda} = \sum_{(X,y)\epsilon N}\sum_{i,j,k}\left\|G_{i,j,k}(y_k) - \emptyset_{i,j,k}(X,\lambda)\right\|^2$$

上式中假设数据$N=\{X,y\}$其中 X 为图像数据，y 为标签数据。λ为权重。

上式中的$G_{i,j,k} = \frac{1}{2\pi\sigma^2}e^{-[(y_k^1-i)^2+(y_k^2-j)^2]/2\sigma^2}$

该损失函数计算的是 CNN 神经网络中输出的热力图与标签中的目标的坐标的高斯分布的距离和。

同理 spatial fusion net 之后的 loss2 也是相同的计算方式。两者是为了保证学习的内容不同。

2.2.4 光流法增强热力图

光流法增强热力图的算法步骤为：

Step1：使用密集光流将附近帧的信号与当前帧对齐。

Step2：使用附加卷积层将这些置信度合并到复合置信图中。

Step3：对每一帧的最终上半身姿势估计就是复合图中最大置信度的位置。

光流法增强热力图的算法流程如下图所示：

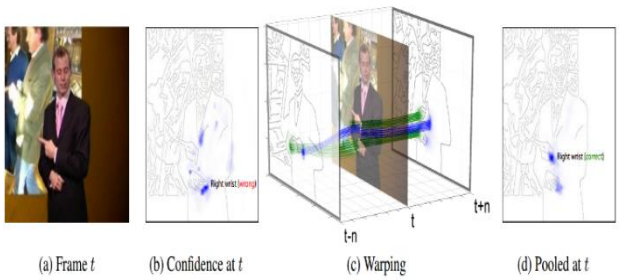


图 3 光强法增强热力图流程图

2.2.5 Flowing ConvNet 方法评价

作者在论文中采用 BBC Pose 数据集进行测试，结果如下图所示：

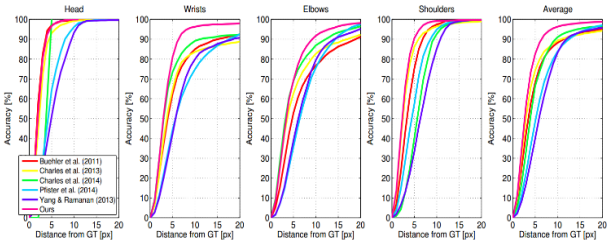


图 4 BBC Pose 测试结果

从上图可以该方法看出，在头部识别的正确率已经接近 100%。而在上半身其他部位的识别率超过 95%。且该算法的正确率明显高于其他方法。

而在 FLIC 数据集中，对于手腕和肘部的 PCK 可以达到 92%，而且可以做到实时性，速度为 5fps。但是该方法对于姿势的估计范围有限。只能对人体上半身的关节点进行识别而不能识别人体下半身的骨骼点。

2. 3 Convolutional Pose Machines方法

2.3.1 方法概述

Convolutional Pose Machines，简称 CPM 方法，是卡耐基梅隆大学机器人研究所的 Shih-En Wei 等人在 2016 年 CVPR 会议中提出的。

该方法使用顺序化的卷积架构来表达空间信息和纹理信息。CPM 方法具有很强的鲁棒性，之后的很多人体姿势识别方法，包括 CMU 的开源项目 OpenPose 都是基于 CPM 方法改进的。

该方法主要有三个创新点：

（1） 用各部件之间的响应图来表达各部件之间的空间约束。响应图和特征图一起作为数据在网络中传递。

（2） 网络分为多个阶段，各个阶段都有监督训练，避免过深网络难以优化的问题。

（3） 使用同一个网络，同时在多个尺度处理输入的特征和响应。既能保证精读，又考虑了各个部件之间的远近距离关系。

输入特征图与各个阶段响应图的示例如下图所示：

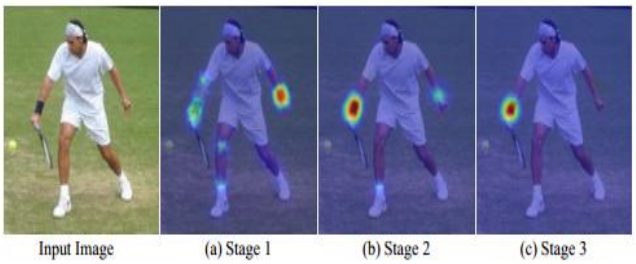


图 5 输入特征图与各个阶段响应图示例图

从上图可以看出以右手为例，第一阶段的特征图识别结果非常不准确，第二阶段大幅改善了第一阶段特征图的结果，但仍然不是人准确。而第三幅图的结果已经非常可靠了。

算法的流程图如下：

Step1：在每一个尺度下，计算各个部件的响应图。

Step2：对于每一个部件，累加所有尺度的响应图，得到总的响应图。

Step3：在每个部件总响应图上，找出总响应最大的点，为该部件的位置。

2.3.2 网络框架

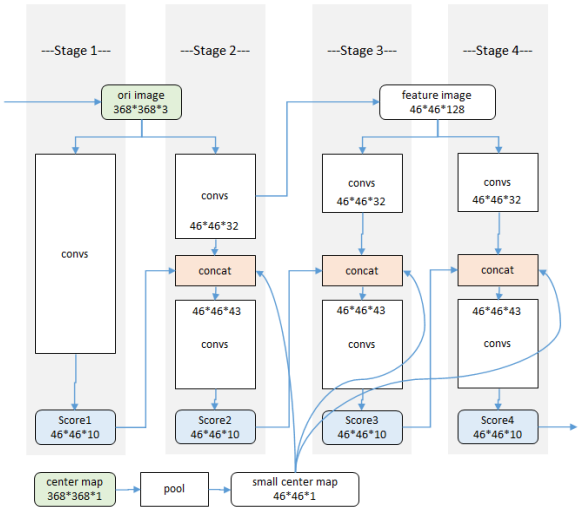


图 6 CPM 方法网络框架

CPM 方法采用的是 cascaded 网络，但 CPM 是一个交互的网络架构，即上一个 FCN 网络上下文会作为下一个 FCN 网络的输入。

网络的输入为彩色图像，共分为四个阶段。每个阶段都能输出各个部件的响应图，使用时以最后一个阶段的响应图为准。

最下方的 center map 是一个提前生成的高斯函数模版，因为 CPM 方法针对的是单人姿态估计问题，如果图片中有多个人，那么 center map 可以告诉网络目前要处理的那个人的位置。

CPM 方法网络各阶段的详细设计如下：

（1） 第一阶段：第一阶段是一个基本的卷积网络，可以从彩色图像直接预测每个部件的响应。半身模型有 9 个部件，另外包含一个背景响应，共 10 层响应图。全身模型则有 15 个部件。

（2）第二阶段：第二阶段也是从彩色图像预测各部件的响应，但是在卷积层中多了一个串联层，用于把反映纹理特征的阶段性卷积结果、反映空间特征

的前一阶段各个部件的相应以及中心约束合一。

（3）第三阶段：第三阶段不再使用原始图像作为输入，而是从第二阶段中途取出一个深度为 128 的特征图作为输入。同样使用串联层综合纹理特征、空间特征以及中心约束这三种因素。

（4）后续阶段：第四阶段结构和第三阶段完全相同，在设计更复杂的网络时，只需要调整部件的数量并重复第三阶段结构即可。

2.3.3 CPM 方法数据训练

（1） 数据扩展：为了丰富训练样本，对原始图片进行随机旋转、缩放、镜像。

（2） 数据标定：姿态数据集中标定的是各个部件的位置，可以通过在每个关键点的真实位置上放置一个高斯响应，来构造响应图的真值。

（3） 中继监督优化：如果直接对整个网络进行梯度下降，输出层的误差经过多层反向传播会大幅减小，即发生 vanishing gradients（梯度消失）现象。如下图所示：

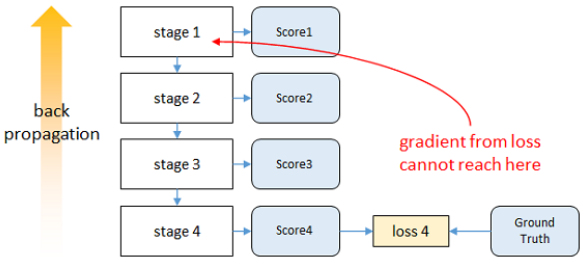


图 7 梯度消失现象示意图

围栏解决此问题，CPM 方法在每个阶段的输出上都计算损失，这样可以保证底层参数正常的进行更新。这种方法称为 intermediate supervision（中继监督法）,中继监督法的示意图如下所示：

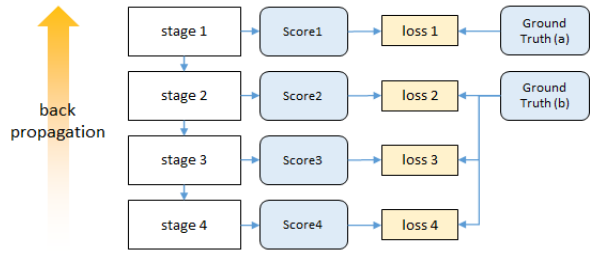


图 8 中继监督法示意图

（4） 多尺度：训练时，已经对数据进行了尺度扩充。在测试时直接从原图像生成不同尺幅的图像，分别送入网络中。最后将所得结果求和。

2.3.4 CPM 方法评价

在 MPII、LSSP 和 FLIC 三个数据集上，CPM 方法的 PCK 指标均超过已有文献，如下图所示：

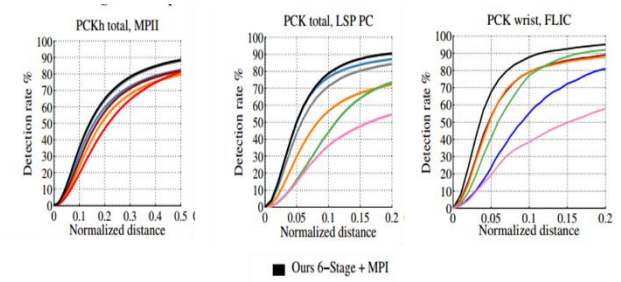


图 9 CPM 方法评价指标

由上图可以看出，虽然与 Flowing ConvNet 相比结果的准确度有所降低，但考虑到 CPM 方法可以对人体全身进行姿态识别，且仍能保持超过 90% 的准确度，效果已经非常不错。

但 CPM 方法仍存在不足，比如在对视频中的人物进行姿态估计时的实时性较差。

2.4 单人姿态识别小结

从图 1 可以看出，截止到目前，最新的单人姿态识别方法已经将准确率提高到了 0.939，单人姿态识别的研究已经基本上趋于饱和。

而且单人姿态识别的应用场景以及应用范围均不如多人姿态识别，单人姿态识别的方法对于检测多人的效果不佳，因此人体姿态识别的方向应该着力于多人姿态识别的研究。

3 多人姿态识别

3.1 多人姿态识别研究现状

3.1.1 多人姿态识别研究方法：

目前，多人姿态识别有两种主流研究方法：

（1）自顶向下（top-down）：这种方法先检测出多个人，再对每一个人进行姿态识别。因此这种方法可以通过目标检测（Object Detection）加上单人姿态估计的方法实现。

（2）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。

3.1.2 多人姿态识别面对的问题：

（1）图像中不知道有多少人，在什么位置。

（2）人与人之间因接触、遮挡而使问题变得复杂。

（3）实时性的要求，图像中的人越多，计算复杂度就越大。

3.1.3 多人姿态识别性能评价指标：

多人姿态识别仍采用 PCK 作为评价指标。

评价多人姿态识别性能好坏的两大数据集为 MPII Multi-Person Dataset 和 MSCOCO Keypoints。目前 MPII Multi-Person Dataset 部分排名如下表所

示：

表 2 MPII Multi-Person Dataset 排名

Method	PCKh	Method	PCKh
Fierau, CVPR18	78.0	Varad., arXiv17	72.2
Newell, NIPS17	77.5	Levink., CVPR17	70.6
Fang, arXiv16	76.7	Insaf., arXiv16	70.0
Cao, CVPR17	75.6	Insafu., ECCV16	59.5
Insaf., CVPR17	74.3	Iqbal, ECCVw16	43.1

3.2 Deepcut&Deepercut方法

3.2.1 方法概述：

Deepcut 方法采用了自顶向下的方法针对多人场景下进行姿态识别。该方法首先采用 CNN 神经网络检测人体，得到人体的候选区域，再判断这些关节属于哪一个人，最后采用整数线性规划（integer linear programming）对模型进行优化。

3.2.2 Deepcut 网络模型：

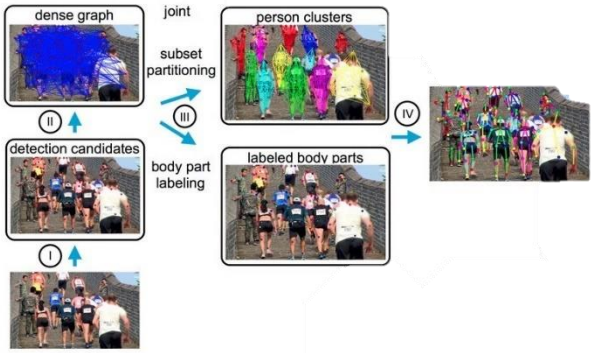


图 10 Deepcut 网络模型图

Deepcut 方法首先使用 fast R-CNN 神经网络对人体进行检测，提取人体的候选区域（body part candidaties）,每一个候选区域对应的是一个关节点，每一个节点作为图中的一个节点。所有的这些候选关节点组成一幅完整的图。节点之间的关联性作为图中的节点之间的权重。这时，可以将其看做是一个优化问题。将属于同一个人的关节点归为一类，每一个人作为一个单独的类。同时，另一条分支需要对检测出来的节点进行标记，确定他们属于人体哪一部分。最后，使用分类的人结合标记的部分构成最终每个人的姿态估计。

3.2.3 Deepcut 方法优点：

（1） 可以在人数和每个人位置未知的情况下解决多人姿态估计问题。通过归类可以得到每个人的关节点分布。

（2） 通过图论节点的聚类问题，有效的使用了非极大抑制问题。

（3） 优化问题可以可以采用整数线性规划（ILP）进行求解，可以采用数学方法得到有效的求解。优化问题的模型为：

$$\arg\min_{(x,y,z)\in X_{DC}} \langle \alpha, x\rangle + \langle \beta, y\rangle$$

$$s.t.\begin{cases} \alpha_{dc} := log\dfrac{1-p_{dc}}{p_{dc}} \\ \beta_{dd'cc'} := log\dfrac{1-p_{dd'cc'}}{p_{dd'cc'}} \\ \langle \alpha, x\rangle := \sum_{d\in D}\sum_{c\in C}\alpha_{dc}x_{dc} \\ \langle \beta, z\rangle := \sum_{dd'\in\binom{D}{2}}\sum_{cc'\in C}\beta_{dd'cc'}z_{dd'cc'} \\ \forall d\in D\forall cc'\in\binom{C}{2}: x_{dc}+x_{dc'}\le 1 \\ \forall dd'\in\binom{D}{2}: y_{dd'}\le\sum_{c\in C}x_{dc} \\ y_{dd'}\le\sum_{c\in C}x_{d'c} \\ \forall dd'd''\in\binom{D}{3}: y_{dd'}+y_{d'd''}-1\le y_{dd'} \\ \forall dd'\in\binom{D}{2}\forall cc'\in C^2: x_{dc}+x_{d'c'}+y_{dd'}-2\le z_{dd'cc'} \\ z_{dd'cc'}\le x_{dc} \\ z_{dd'cc'}\le x_{d'c'} \\ z_{dd'cc'}\le y_{dd'} \\ \forall dd'\in\binom{D}{2}\forall cc'\in C^2: x_{dc}+x_{d'c'}-1\le y_{dd'} \end{cases}$$

3.2.4 Deepcut 方法不足

由于使用了自适应的 fast R-CNN 进行人体的检测，同时又使用 ILP 进行人体姿态估计。所以该方法的计算复杂度非常大。

3.2.5 Deepcut 方法改进

考虑到 Deepcut 存在计算复杂度非常高的缺点，Deepercut 方法在 Deepcut 的基础上，对其进行改进，改进的方式基于以下两个方面：

（1） 使用最新提出的 Residual Net 对人体进行检测，与 Deepcut 方法使用的 fast R-CNN 网络相比效果更佳精确，精度更高。

（2） 提出 Image-Conditioned Pairwise Terms 方法，能够将众多候选区域的节点压缩到更少数量的节点，该方法的原理是通过候选节点之间的距离来判断其是否为同一个重要节点。Image-Conditioned Pairwise Terms 的描述如下图所示：

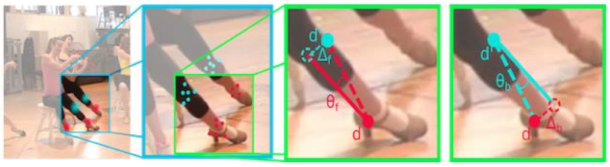
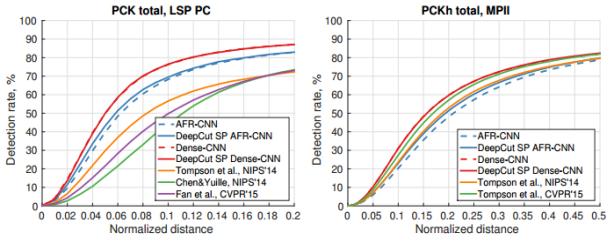


图 11 Image-Conditioned Pairwise Terms 方法

3.2.6 Deepcut&Deepercut 方法评价



(a) LSP (PC) (b) MPII Single Person

图 12 Deepcut 方法在 LSP 和 MPII 单人数据集结果

从上图可以看出，Deepcut 方法对于单人姿态估计，在 LSP 数据集上的 PCK 达到 87%左右，在 MPII 数据集上的 PCK 达到 82.5%左右。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存在差距。

图 13 Deepcut 方法在 MPII Multi-Person 数据集结果

Unary	Pairwise	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP	time [s/frame]
DeepCut [10]	*DeepCut* [10]	50.1	44.1	33.5	26.5	33.0	28.5	14.4	33.3	259220
DeepCut [10]	this work	68.3	58.3	47.4	38.9	45.2	41.8	31.2	47.7	1987
this work	this work	**70.9**	59.8	**53.1**	**44.4**	50.0	46.4	39.5	52.3	1171
+ location refinement before NMS		70.3	**61.6**	52.1	43.7	**50.6**	**47.0**	**40.6**	**52.6**	**578**

对于多人姿态估计，在 MPII 多人数据集上的 AP 达到 60.5%，但识别的速度非常慢。

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	UBody	FBody
AFR-CNN det ROI	71.1	65.8	49.8	34.0	47.7	36.6	20.6	55.2	47.1
AFR-CNN MP	71.8	67.8	54.9	38.1	52.0	41.2	30.4	58.2	51.4
AFR-CNN MP UB	75.2	71.0	56.4	39.6	-	-	-	60.5	-
Dense-CNN det ROI	77.2	71.8	55.9	42.1	53.8	39.9	27.4	61.8	53.2
Dense-CNN MP	73.4	71.8	57.9	39.9	**56.7**	**44.0**	**32.0**	60.7	**54.1**
Dense-CNN MP UB	**81.5**	**77.3**	**65.8**	**50.0**	-	-	-	**68.7**	-
AFR-CNN GT ROI	73.2	66.5	54.6	42.3	50.1	44.3	37.8	59.1	53.1
Dense-CNN GT ROI	78.1	74.1	62.2	52.0	56.9	48.7	46.1	66.6	60.2
Chen&Yuille SP GT ROI	65.0	34.2	22.0	15.7	19.2	15.8	14.2	34.2	27.1

Table 6. Pose estimation results (AP) on MPII Multi-Person.

图 14 Deepercut 方法在 MPII Multi-Person 数据集结果

Deeper 方法在 MPII Multi-Person 数据集中的测试结果与 Deepcut 方法相比，准确率和速度都有所提升，尤其是速度方面，提高了 221 倍！

3. 3 OpenPose方法

3.3.1 方法概述：

Deepcut 方法采用的是自上而下的人体姿态估计算法，这种方法的运算时间会随着图像中人的个数的增加而显著增加。而 OpenPose 采用的是自下而上的人体姿态估计算法，先得到关键点的位置再获得骨架，因此这种方法所需的计算时间基本不变。

OpenPose 的前身就是 CPM 方法，但相较于 CPM 方法只能识别单人，OpenPose 方法提出将人体部位与个体联系起来的非参数化方法—PAFs 更加优雅地解决了多人姿态识别问题，而且在性能和效率上均得到了非常不错的结果。

3.3.2 网络架构：

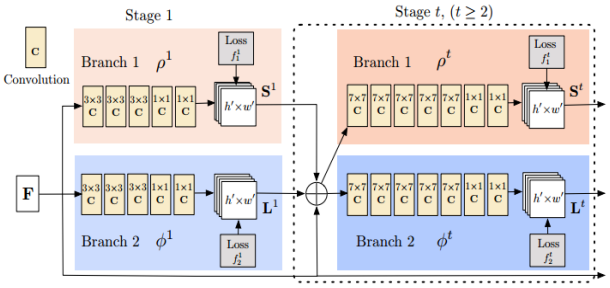


图 15 OpenPose 方法网络架构

OpenPose 方法使用了反复迭代的 CNN 网络对人体姿态进行检测，每个 CNN 网络都有两个分支，第一个分支用于计算部位置信图（Part Confidence Map），即关节点；第二个分支用于计算部分亲和域（Part Affinity Fields），即肢体躯干。

从图中还可以看出，网络的第 1 个阶段接受的输入是特征 F，经过网络的处理后分别得到 S^1 和 L^1，从第 2 阶段开始，阶段 t 的网络的输入包括三部分：S^{t-1}、L^{t-1}、F。

这样反复迭代，直到网络变得收敛。

3.3.3 损失函数

OpenPose 方法为每一阶段的两个分支网络分别设计损失函数，对损失进行空间加权，来解决一些数据集不能完全标记所有人的问题。

损失函数如下所示：

$$f_S^t = \sum_{j=1}^{J} \sum_{p} W(p) \cdot \left\| S_j^t(p) - s_j^*(p) \right\|_2^2$$

$$f_L^t = \sum_{c=1}^{C} \sum_{p} W(p) \cdot \left\| L_c^t(p) - L_c^*(p) \right\|_2^2$$

其中 W 为 0-1 变量，当图像位置 p 没有注释时，$W(p) = 0$，避免训练时惩罚 TP 预测。

GT 置信图根据标注的人体关键点位生成，以多人为例，第 k 个人的第 j 个可见部位的置信图表示为$S_{j,k}^*$：

$$S_{j,k}^*(p) = \exp\left(-\frac{\left\| p - x_{j,k} \right\|_2^2}{\sigma^2}\right)$$

其中$p \in \mathbb{R}^2$，$x_{j,k}$为关键点标注位置。则：

$$S_j^*(p) = \max_k S_{j,k}^*(p)$$

当点 p 在 k 的肢体 c 上时，$L_{c,k}^*(p)$等于从关键点 j_1 指向关键点 j_2 的单位向量。否则为零向量。如下图所示：

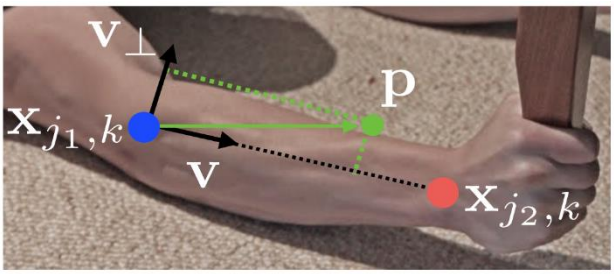


图 16 OpenPose 亲和场图

公式为：

$$L_{c,k}^*(p) = \begin{cases} v & p在k的肢体c上 \\ 0 & p不在k的肢体c上 \end{cases}$$

肢体上的点必须满足以下以下条件：

$$0 \le v \cdot (p - x_{j1,k}) \le l_{c,k} and \left| v_\perp \cdot (p - x_{j1,k} \right| \le \sigma_l$$

其中肢体长度：$l_{c,k} = \left\| x_{j2,k} - x_{j1,k} \right\|_2$

肢体宽度：σ_l为像素距离。

最终 PAFS 为图像中所有人的平均亲和场：

$$L_c^*(p) = \frac{1}{n_c(p)} \sum_k L_{c,k}^*(p)$$

其中，$n_c(p)$为点 p 处所有人产生的非零向量个数。

测试阶段，通过沿着连接候选部位位置的线段，计算相应 PAF 上的线积分来测量候选部位检测之间的关联，公式如下：

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot \frac{d_{j2} - d_{j1}}{\left\| d_{j2} - d_{j1} \right\|_2} du$$

$p(u)$为两个人体部位 d_{j1} 和 d_{j2} 的位置插值。

$$p(u) = (1-u)d_{j1} + ud_{j2}$$

每个阶段的中间监督通过定期补充梯度来解决梯度消失问题。总损失为：

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t)$$

3.3.4 PAFs 多人分析

通过对检测置信度图执行非最大值抑制，以获得一组离散的部位候选位置。每个部位可能产生多个候选，经组合生成大量的候选肢体。通过 PAF 上计算线积分来对每个候选肢体打分。找到最优解的问题对应于一个已知为 NP 难的 K 纬匹配问题。

该问题中，结点为 D_{j1} 和 D_{j2} 边是两组结点间所有可能的连接。每条边的权重为之前计算的线积分 E。二部图中的匹配是以没有两条边共享一个结点的方式选择的边的子集。优化的目标是为选定的边找到最大权重的匹配。优化模型如下所示：

$$\max_{Z_c} E_c = \max_{Z_c} \sum_{m \in D_{j1}} \sum_{n \in D_{j2}} E_{mn} \cdot z_{j1j2}^{mn}$$

$$s.t. \begin{cases} \forall m \in D_{j1}, & \sum_{n \in D_{j2}} z_{j1j2}^{mn} \le 1 \\ \forall n \in D_{j2}, & \sum_{m \in D_{j1}} z_{j1j2}^{mn} \le 1 \end{cases}$$

通过匈牙利算法即可获取最优匹配。

3.3.5 OpenPose 方法评价

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	**mAP**	s/image
Fig. 6b	91.8	**90.8**	80.6	69.5	78.9	71.4	63.8	78.3	362
Fig. 6c	92.2	90.8	80.2	69.2	78.5	70.7	62.6	77.6	43
Fig. 6d	92.0	90.7	80.0	69.4	78.4	70.1	62.3	77.4	0.005
Fig. 6d (sep)	**92.4**	90.4	**80.9**	**70.8**	**79.5**	**73.1**	**66.5**	**79.1**	**0.005**

图 17 OpenPose 在 MPII Single-Person 测试集的结果

Team	**AP**	AP^{50}	AP^{75}	AP^M	AP^L
Test-challenge					
Ours	**60.5**	**83.4**	**66.4**	55.1	**68.1**
G-RMI [19]	59.8	81.0	65.1	**56.7**	66.7
DL-61	53.3	75.1	48.5	55.5	54.8
R4D	49.7	74.3	54.5	45.6	55.6
Test-dev					
Ours	**61.8**	**84.9**	**67.5**	57.1	**68.2**
G-RMI [19]	60.5	82.2	66.2	**57.6**	66.6
DL-61	54.4	75.3	50.9	58.3	54.3
R4D	51.4	75.0	55.9	47.4	56.7

图 18 OpenPose 在 MPII Multi-Person 测试集 的结果

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	**mAP**	s/image
Subset of 288 images as in [22]									
Deepcut [22]	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
Iqbal et al. [12]	70.0	65.2	56.4	46.1	52.7	47.9	44.5	54.7	10
DeeperCut [11]	87.9	84.0	71.9	63.9	68.8	63.8	58.1	71.2	230
Ours	**93.7**	**91.4**	**81.4**	**72.5**	**77.7**	**73.0**	**68.1**	**79.7**	**0.005**
Full testing set									
DeeperCut [11]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Iqbal et al. [12]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Ours (one scale)	89.0	84.9	74.9	64.2	71.0	65.6	58.1	72.5	0.005
Ours	**91.2**	**87.6**	**77.7**	**66.8**	**75.4**	**68.9**	**61.7**	**75.6**	**0.005**

图 19 OpenPose 在 COCO 测试集的结果

从上图中可以看出，无论是在 MUII 单人测试集还是多人测试集或者是 COOC 测试集中，OpenPose 在性能和准确率上均取得了非常优秀的结果。

4 总结

4.1 人体姿态识别存在的问题

4.1.1 单人姿态识别存在的问题

目前单人姿态识别的准确率已经很高，在 MPII Single-Person 测试集上最高的准确率已经达到 93.9%，相关的研究也已经趋于饱和。但岁从目前来看，尽管采用不同的神经网络和不同的方法，识别的准确率的提升都非常有限了，因此如何将单人姿态识别的准确率进一步提升是当前单人姿态识别面对的主要问题。

4.1.2 多人姿态识别存在的问题

（1） 对于自顶向下的方法来说，检测的误差是一个很大的影响。而且即使在检测任务是正确的情况下，提取的信息也不适用与单人的姿态估计方法。同时，冗余的检测框也使得单人的姿态被重复估计。

（2） 对于自底向上方法来说，当两个人比较靠近时，人体关键点分配到每个人身上时会出现偏差，这是该方法面临的最大问题。

（3） 目前大多数的人体姿态识别是 2D 姿态识别，但仅仅有二维的姿态不能完全反映出人体的姿态信息。

（4） 目前多人姿态识别的速度仍然不够快，性能较差。

4.1.3 多人姿态识别存在的问题一些解决思路

（1） 针对 4.1.2 中的问题（1）和问题（2），上海交通大学的卢策吾团队提出了一个 RMPE 框架，采用自顶向下的方法，进行姿态估计。取得了一定的进展。

（2）针对 4.1.2 中的问题（3），Facebook AI Reaserch（FAIR）构建了一个 DensePose-RCNN 系统，将 2D RGB 图像的所有人类像素实时映射至 3D 模型。

4.2 人体姿态识别未来的发展趋势

通过对文献中相关技术的学习，我们认为人体姿态识别未来主要的发展方向主要有三个：

（1） 将单人姿态识别与多人姿态识别相结合，即一个系统实现对单人和多人姿态的识别。

（2）对 3D 的人体姿态直接估计，而非将 2D 的人体姿态映射至 3D 模型之上。

（3）对识别出的人体姿态进行分析，理解不同人体姿态所表现出的信息。

参 考 文 献

[1] Zhou Yi-Nong, Human gesture recognition in fixed scenes. Computer Programming Skills & Maintenance, 2018, 57(11): 150-151 (in Chinese) (赵一侬.固定场景下人体姿态识别. 电脑编程技巧与维护, 2018, 57(11): 150-151)

[2] Tomas Pfister, James Charles, Andrew Zisserman. Flowing convnets for human pose estimation in videos //Proceedings of the ICCV. Santiago, Chile, 2015: 1-8

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E.Hinton. Imagenet clas sifaction with deep convolutional neural networks // Proceedings of the NIPS. Lake Tahoe, America, 2012: 2-5

[4] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh. C onvolutional pose macines // Proceedings of the CVPR. Las Vega s, America, 2016: 1-8

[5] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele. Deepcut:joint su bset partition and labeling for multi person pose estimation //Proc eedings of the CVPR. Las Vegas, America, 2016: 3-7

[6] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo A ndroluka, Bernt Schiele. Deepercut:a deeper,stronger,and faster mul ti-person pose estimation model //Proceedings of the ECCV. Amst erdam, The Netherlands, 2016: 7-14

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep resid ual learning for image recognition //Proceedings of the CVPR. La s Vegas, America, 2016: 1-8

[8] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh. Realtime mul ti-person 2d pose estimation using part affinity fields //Proceeding s of the CVPR. Hawaii State, America, 2017: 1-7

[9] Karen Simonyan, Andrew Zisserman. Very deep convolutional net works for large-scale image recognition //Proceedings of the ICL R. San Diego, America, 2015: 3-12

[10] Haoshun Fang, Shuqin Xie, Cewu Lu. Rmpe:regional multi-perso n pose estimation //Proceedings of the ICCV. Venice, Italy, 2017: 4-8

[11] Riza Alp Guler, Natalia Neverova, Iasonas Kokkinos. Densepose: dense human pose estimation in the wild. arXiv, 2018: 3-10