基于多分支 Faster RCNN 的人体姿态估计

魏凯强，赵旭，王磊
上海交通大学自动化系，上海，200240

摘要：人体姿态估计是计算机视觉领域备受关注的研究热点之一。一般姿态估计方法主要针对单人图像进行姿态估计，本文提出多分支的 Faster RCNN 网络结构可以同时检测图片中多个人及其肢体部件，并在此基础上估计人体姿态。本文的多分支结构模型一方面能够提高人体和人体部件的检测效果，另一方面可以利用共享卷积加速检测过程。在多个公开数据集上的对比实验证实了多分支结构模型比原始模型在检测人体及人体部件的任务上效果更好，与其他的姿态估计算法相比，本文方法也达到了相当或更优的性能指标。

关键词：卷积神经网络；多分支 Faster RCNN；部件检测；姿态估计

Multi-branch Faster RCNN for Human Pose Estimation

Wei Kaiqiang, Zhao Xu, Wang Lei
Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240

Abstract: Human pose estimation has long been a deeply concerned research topic in computer vision field. Most human pose estimation algorithms only focus on images with single person. We propose a multi-branch Faster RCNN model to detect multiple people as well as their parts in the image and then get their poses. Our multi-branch model can improve performance on detecting human parts and persons, while speed up detection process with sharing weights. Experiments on several public datasets demonstrate that our multi-branch model outperforms the original model on person and parts detection, and achieves fairly or better results when compared with other pose estimation algorithms.

Key words: convolution neural network; multi-task Faster RCNN; parts detection; pose estimation

1 引言

人体姿态估计是计算机视觉领域的研究热点之一。快速鲁棒的姿态估计有着广泛的应用，如人机交互，虚拟现实，智能监控等。人体姿态估计的任务是定位图片中人体各个关节点位置，并有序的连接关节点得到人体姿态结构信息。由于人的动作变化复杂多样，各个部件之间以及不同人之间会相互遮挡，背景物体的干扰等因素的影响，使得人体姿态估计成为一项极具挑战的任务。

PS(Pictorial Structure)[1]是典型的人体结构建模方法，在姿态估计中应用广泛。PS 模型用矩形框表示人体各部件的位置，用树形结构表征各部件之间的连接关系。Yang[2]在 PS 模型基础上引入“类型”的概念，利用多个混合类型的模板来建模人体的各个部件，提高了姿态估计准确率。这类方法主要针对单人姿态估计问题。

近年来，卷积神经网络（Convolution Neural Network）[3]在基于图像的物体分类和检测任务上取得巨大进展，受到广泛关注。研究者也将卷积神经网络用于解决人体姿态估计和动作识别等问题。Deeppose[4]训练了三阶段级联的卷积神经网络，利用回归方法较为准确地预测人体关节点位置，证实了卷积神经网络在姿态估计问题上的有效性。Gkioxari[5]用卷积神经网络训练出人的部件检测器用于人的动作和属性识别。

基金项目：国家自然科学基金（61273285,61375019）
第一作者简介：魏凯强，男，硕士研究生在读，研究方向为计算机视觉，深度学习，邮箱：wakerkq@sjtu.edu.cn

本文从检测人体部件入手，根据部件检测结果定位关节点位置，估计人体姿态。在Faster RCNN[6]物体检测算法的基础上，本文提出多分支 Faster RCNN 网络结构用于同时检测图片中的多个人及其部件，并利用共享卷积加速检测过程。在检测出部件和人的位置后，根据人体布局先验，制定一套规则定位人体关节点位置，估计人体姿态信息。在多个公开数据集上的对比实验说明了本文多分支模型比原始模型在检测人体及人体部件的任务上效果更好，与其他姿态估计方法相比，本文方法也达到了相当或更优的性能指标。

2 多分支 Faster RCNN 模型

2.1 Faster RCNN 原理

Faster RCNN[6]在 Fast RCNN[7]基础上提出利用 RPN (Region Proposal Network)产生高质量的物体候选框，再由 Fast RCNN 进行检测，取得了较好的效果。RPN 结构是在基础卷积网络的顶层构建两层卷积层，构成全卷积网络，其输出有两项，一项是窗口是否包含物体的分类信息，另一项是窗口的位置信息。RPN 采用公式(1)所示的多任务损失函数训练网络。

$$L(p_i,t_i)=L_{cls}(p_i,p_i^*)+\lambda p_i^* L_{reg}(t_i,t_i^*) \quad (1)$$

其中，$L_{cls}(p_i,p_i^*)$表示窗口i是否包含物体的分类损失函数，p为预测值，p^*为真值，窗口包含物体时真值为 1，否则为 0，λ为平衡系数。$L_{reg}(t_i,t_i^*)$使用[7]定义的回归损失函数，如公式(2)所示，该损失函数使得模型训练更容易收敛。

$$L_{reg}(t_i,t_i^*)=\begin{cases}0.5(t_i-t_i^*)^2, & |t_i-t_i^*|\le 0.5\\ |t_i-t_i^*|-0.5, & |t_i-t_i^*|>0.5\end{cases} \quad (2)$$

其中，$t_i=\{t_x,t_y,t_w,t_h\}$表示预测框位置参数，$t_i^*=\{t_x^*,t_y^*,t_w^*,t_h^*\}$表示真值框位置参数。

2.2 多分支模型结构框架

为了同时检测图片中人和人的各个部件，本文提出多分支的 Faster RCNN 型，其结构框架如图 1 所示。与原始模型不同，本文模型的网络顶层具有两个分支结构：部件分支和人体分支。部件分支用于检测人的六种部件，包括头，躯干，上臂，前臂，大腿，小腿；人体分支用于检测人。两个分支

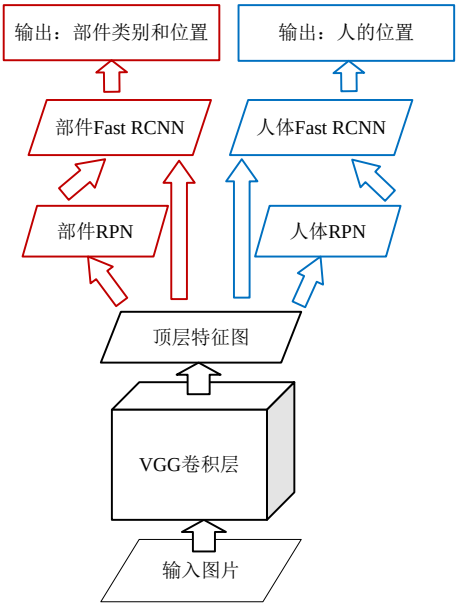

图 1 多分支模型结构框架

结构共享底层的 VGG[8]卷积层。每个顶层分支和底层的卷积网络组成一个完整的 Faster RCNN 模型。

构建上述具有两个分支结构的 Faster RCNN 模型有三个原因。第一，本文把人体和部件分开检测是由于人和部件不是相互独立的，而是整体和部分的关系，人的位置框会包含部件的位置框。如果把人和部件放到一个模型中，则会影响 RPN 训练产生的物体候选框的质量，进而影响整个模型的检测效果。本文通过对比实验证实了这一想法，详见实验部分。第二，为提高测试速度，本文实现了人体检测模型和部件检测模型共享底部的卷积层。测试阶段，前向推理过程计算量主要集中在底部多层卷积结构。两个顶层分支共享底部卷积层参数，则只需进行一次前向过程，加速了测试过程。第三，由于人的检测结果较为准确可靠，可用来排除把背景物体误检为部件的情况，有助于提高部件检测效果。

2.3 模型训练

为实现检测人的模型和检测部件的模型共享底部卷积层，本文借鉴文献[6]，提出三阶段训练方法，训练过程详述如下：

第一阶段先训练初始的部件检测模型，即训练 RPN 模型产生部件的候选框，并给到 Fast RCNN 中得到部件检测模型。该阶段的 RPN 和 Fast RCNN 网络都用 ImageNet 数据库预训练的模型参数来初始化。

第二阶段重新训练部件检测模型。该阶段 RPN 和 Fast RCNN 模型初始化使用第一阶段训练的 Fast RCNN 模型参数，并且固定 RPN 和 Fast RCNN 底部共享的卷积层参数不变，只微调它们各自独有的顶层分支。此时已训练出完整的部件检测模型。

第三阶段训练人的检测模型。先训练人的 RPN 模型，使用第一阶段的部件 Fast RCNN 模型参数初始化模型，并固定底层的卷积层参数不变，只微调 RPN 顶层的两层卷积结构。再训练人的 Fast RCNN 模型，同样使用第一阶段部件的 Fast RCNN 模型参数初始化模型，固定底层的卷积层参数不变，只微调 Fast RCNN 顶层独有的结构。

至此，人体检测模型和部件检测模型共享了底层的 VGG 卷积层参数，得到多分支检测模型。测试阶段，只需对底层共享的卷积层进行一次计算过程，并把得到的特征给到部件分支和人体分支，就可以同时得到部件和人的检测结果。

2.4 姿态估计

姿态估计需要定位各个关节点，并将其连接起来得到姿态信息。本文根据部件位置来定位关节点。为把关节点定位转换为部件检测，并在得到部件位置后还原出关节点位置，本文制定了一套规则，在保证精度的前提下能够迅速定位关节点。一般图片中可能有多个人，估计多人姿态还需要把不同人的关节点与相应的人匹配起来。

2.4.1 关节点定位和部件检测的转换

训练阶段，现有数据库标注的仅是关节点位置，本文用一个部件的两个关节点构建部件位置矩形框，并对得到的矩形框扩大 1.2 倍，使得矩形框能够完整包括部件。由此可训练出部件检测模型。

部件检测模型的输出是部件的位置矩形框。为检测出关节点的位置，先把部件矩形框缩小 1/1.2 倍。当矩形框的长宽比小于 0.3 时，则取短边的中点为关节点位置；当长宽比大于 0.3 时，则取出矩形框内的图像块，判断其中包含的部件朝向，计算其沿矩形两个对角线方向的梯度，把梯度大的方向上两个对角点作为关节点位置。

2.4.2 姿态的连接

定位各部件的关节点位置后，需要根据人体布局将其有序连接起来。按照距离关系匹配上臂和前臂，并把肘关节点位置修正为上臂和前臂对应的肘关节点两个位置的中心。同理匹配大腿和小腿，并把膝关节点位置修正为大腿和小腿对应的膝关节点两个位置的中心。

多分支模型能同时检测出所有人和人的部件，可根据人的位置框，找到与其有重叠的部件作为该人的部件。人的检测结果较可靠，可用来排除一些把背景物体当成部件的误检。依次处理每个人的各部件即可得到多人姿态估计结果。

3 实验结果与分析

3.1 多分支模型检测结果

本文使用 MPII 人体姿态数据库[9]训练多分支结构网络模型。MPII 数据库有约 2.5 万张标注了多人关节点的图片。从 MPII 数据库训练集中选取标注完整的图片约 1.7 万张，并按照 4:1 划分训练集和测试集。

实验中，按照 2.4 节所述由关节点标注构建部件的真值位置框，训练部件检测模型。增加 RPN 网络训练阶段和测试阶段产生的物体候选框数目会提高检测结果，如表 1 中所示，表中 AP(Average Precision)即平均精度，是衡量物体检测结果的一种通用指标。测试阶段，人检测的得分阈值设为 0.75，部件的得分阈值设为 0.6。对大于得分阈值的检测框做非最大抑制，把部件的重叠阈值设为 0.4，滤除重叠较大的检测框。

表 1 不同 RPN 候选框个数的检测 AP(%)值

人体部件	训练 2k，测试 300	训练 6k，测试 1k
头	71.8	75.1
躯干	69.0	71.3
上臂	30.1	32.5
前臂	17.7	18.7
大腿	29.9	31.7
小腿	32.3	34.4
人	89.1	89.0

为了对比多分支模型和原始模型[6]，本文按照文献[6]中方法训练出同时检测人体部件和人的模型。对比实验中，两者训练 RPN 阶段产生的候选框个数都设为 2k，测

试候选框个数都设为 1k。结果对比如表 2 中第一列和第二列所示，在 MPII 数据集上部件和人的检测 AP 值都有了较明显的提高。头部检测的 AP 曲线对比如图 2，多分支结构 AP 值为 74.3%，而原始模型的 AP 值仅为 54.7%。本文多分支结构能达到较高的 AP 主要在于召回率更高，说明其 RPN 网络产生的候选框的质量更高。

为了验证模型的泛化能力，本文用在 MPII 数据库上训练的模型在 Pascal VOC09 人体数据库[10]的测试集上进行测试。该测试集有 1446 张测试图片，其中人的尺度变化较大，为检测带来更大挑战。对比表 2 中第三列和第四列中各部件的检测 AP 值，同样验证了本文多分支模型比原始模型的检测结果更好。

表 2 本文模型和原始模型检测 AP(%)值对比

人体部件	MPII 数据库		Pascal 数据库	
	本文模型	原始模型[6]	本文模型	原始模型[6]
头	**74.3**	54.7	**39.7**	32.6
躯干	**70.2**	56.2	**16.5**	13.0
上臂	**30.5**	22.5	**17.8**	14.6
前臂	**18.1**	13.3	**9.6**	7.7
大腿	**30.3**	20.0	**14.5**	8.6
小腿	**32.4**	23.8	**17.9**	11.9
人	**89.2**	85.9	**56.0**	54.2

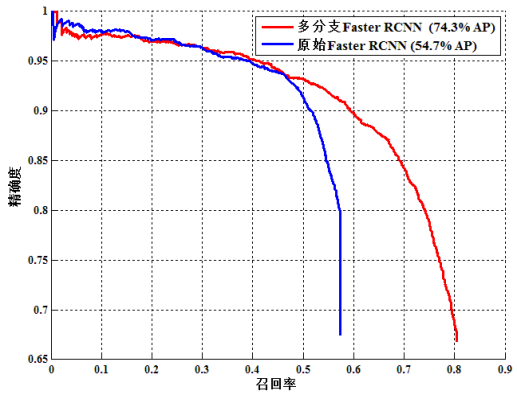

图 2 头的检测 AP 曲线对比

文献[5]用卷积神经网络在 Pascal VOC09 人体数据库上检测人体部件，包括人的头，躯干，腿。本文与其检测结果进行对比如表 3 所示，表中每列表示测试 AP 不同的面积重叠阈值，阈值越高测试指标越严格，由表可知，在高阈值下，本文方法检测部件的 AP 值高于文献[5]的结果，验证了本文模型良好的泛化能力。

表 3 不同阈值下部件检测 AP(%)值对比

阈值		0.3	0.4	0.5
本文模型	头	49.3	45.9	39.8
	躯干	41.9	31.2	17.5
	大腿	26.4	21.4	15.5
	小腿	29.6	24.8	17.5
文献[5]	头	51.8	45.2	31.6
	躯干	36.3	23.6	9.4
	腿	27.9	20.0	10.4

3.2 姿态估计结果

为评估姿态估计方法的效果，本文在 LSP 数据库[11]上进行实验，并与其它姿态估计方法作比较。LSP 及其扩充数据库共包含 12000 张图片，其中训练集 11000 张，测试集 1000 张。对训练图片做水平翻转以扩充数据。实验中，由于该数据库图片分辨率较小，模型训练时 RPN 的窗口的三个尺度参数设置为[16，32，64]，以适应人体部件的大小。模型训练完成后，按照 2.4 节所述规则得到人的姿态，并采用 PCP (Percentage of Correct Parts)指标评估结果。PCP 指标是当一个部件对应的两个关节点的预测位置与真值位置的偏差距离都在部件长度的一半范围内时，认为该部件被正确检测。

姿态估计结果示例如图 3 所示，其中第一行图片来自 MPII 数据库，第二行图片来自 LSP 数据库。本文姿态估计方法在 K40 显卡上测试速度为每张图片约 350ms。在 LSP 数据库上与其他方法对比 PCP 指标如表 4，本文方法估计头部的指标为 87.4，下腿的指标为 74.2，平均指标为 63.0，高于其它方法，验证了本文方法的有效性。

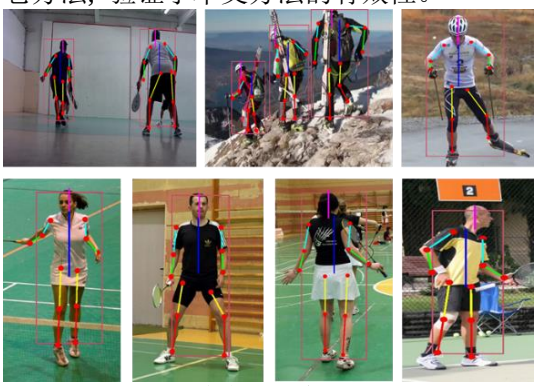

图 3 姿态估计结果示例

表 4　LSP 数据库上姿态估计 PCP 指标对比

方法	头	躯干	上臂	前臂	大腿	小腿	平均值
Andriluka[1]	74.9	80.9	46.5	26.4	67.1	60.7	55.7
Dontone[12]	79.2	81.6	45.1	24.7	66.5	61.0	55.5
Yang[2]	79.3	82.9	**56.0**	**39.8**	70.3	67.0	62.8
Pishchulin[13]	78.1	**87.5**	54.2	33.9	**75.7**	68.0	62.9
本文方法	**87.4**	84.9	47.5	34.6	72.6	**74.2**	**63.0**

4 结论

本文提出的多分支 Faster RCNN 结构比原始结构在检测人和部件的任务上取得了更好的结果，并且利用人体分支和部件分支共享卷积加速了检测过程。该多分支结构具有一般性，同样适用于其他物体和物体部件的检测。在检测结果的基础上，通过制定一套规则估计图片中单人或多人的姿态，与其它姿态估计方法相比，本文方法取得了较好的效果。

参考文献

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, “Pictorial structures revisited: People detection and articulated pose estimation”. In CVPR 2009, pp.1014–1021.

[2] Yang Y., and Deva R. "Articulated human detection with flexible mixtures of parts." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2878-2890.

[3] Krizhevsky A, Sutskever I, and Hinton G E. “Imagenet classification with deep convolutional neural networks” Advances in neural information processing systems. 2012, pp. 1097-1105.

[4] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." In CVPR 2014, pp. 1653-1660.

[5] Gkioxari, Georgia, Ross Girshick, and Jitendra Malik. "Actions and attributes from wholes and parts." Proceedings of the IEEE International Conference on Computer Vision. 2015, pp.2470-2478.

[6] Ren, S., He, K., Girshick, R., & Sun, J. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In Advances in neural information processing systems” 2015, pp. 91-99.

[7] Girshick R. “Fast r-cnn”. Proceedings of the IEEE International Conference on Computer Vision. 2015, pp.1440-1448.

[8] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[9] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and state of the art analysis." In CVPR, 2014, pp. 3686–3693.

[10] Everingham M, Van Gool L, Williams C.K.I, Winn J, and Zisserman A, “The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results, “http://www.pascalnetwork.org/challenges/VOC/voc2009/workshop/index.html”.

[11] Sam Johnson and Mark Everingham, “Clustered pose and nonlinear appearance models for human pose estimation.” in BMVC, 2010, vol. 2, p. 5.

[12] Dantone M, Gall J, Leistner C, et al. “Human pose estimation using body parts dependent joint regressors”. In CVPR 2013, pp 3041-3048.

[13] Pishchulin L, Andriluka M, Gehler P, et al. “Poselet conditioned pictorial structures”. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013, pp 588-595.