基于图模型的姿态分割估计方法

王国桢,方贤勇
(安徽大学 媒体计算研究所,安徽 合肥 230601)

摘 要:计算机视觉领域中现在有一个非常热门的问题就是人体的姿态估计,它可用于行人检测、人体活动分析、人机交互以及视频监控等方面。目前对于图像的人体姿态的估计方法在处理较复杂的背景的时候难以得到理想的效果,其原因在于这些方法不好区分人体和复杂背景,从而无法得到其想要的特征值供其使用。针对这一不足,提出一种姿态分割估计方法。该方法将人体分割后去除复杂背景的影响,并且在图结构模型中,结合使用形状上下文特征的方法进行训练对比,求解得出最优的人体姿态。实验结果表明,该方法可以较好地在复杂背景下获得人体的姿态估计,更好地克服背景带来的干扰,得到较现有方法更加理想的人体估计结果,从而把人体的姿态从复杂的背景图像中给成功地估计出来。

关键词:人体姿态;图结构模型;形状上下文;分割
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2016)12-0053-05
doi:10.3969/j.issn.1673-629X.2016.12.012

Pose Segmentation and Estimation Based on Pictorial Structure Model

WANG Guo-zhen, FANG Xian-yong
(Institute of Media Computing, Anhui University, Hefei 230601, China)

Abstract: Human pose estimation is one of the hot topics in the field of computer vision, and can be used for pedestrian detection, human activity analysis, human-computer interaction and video surveillance and so on. It is difficult to robustly estimate the human pose under the complex background for existing estimation methods of human pose, which is partially due to the lack of good features to separate the foreground human from the complex background. Aiming at the deficiencies mentioned above, a pose segmentation and estimation method is presented. The human is segmented from the background by semantic segmentation. Then shape context method is adopted to obtain the optimal human pose in the pictorial structure. Experimental results show that the proposed method can get the pose estimation, overcome the interference from background, and obtain a better body estimation than the existing method under complex backgrounds. So it can be success to estimate the body pose from the image in a complex background.

Key words: human pose; pictorial structure; shape context; segmentation

1 概 述

当今,在静态图片中对人体的姿态估计成为一个热门话题,在许多人机交互[1-2]自动化的检测、运动、动作识别、角色动画、临床步态分析中精准的人体姿态估计得到了广泛应用。尽管已有多年的研究历史,几个因素使其成为一个极具挑战的内容,图片中的人可以以各种各样的姿态出现。文中针对这一问题进行研究,提出基于图模型的姿态估计方法。

图形结构(Pictorial Structure)[3]模型,是一个可以很好解决这个问题的方法,把人体的各个模块放在一个无向图中来解决,然后可以利用图模型中具有的推理方法估计出人体的姿态。然而这个问题的难度在于,对于这个模型的建立如何有效来表现这些结构的多样性、可变性,并且如何获取能够让这些模块联系在一起的关系结构。Benjamin 等[4]在特征提取中加入 HOF(Histogram of Optical Flower,光流直方图)[5],对于连续图片中人的姿态进行估计有更好的效果。Ouyang Wanli 等[6]是在图形结构模型的基础上,对图片的训练过程使用 Deep Learning[7]的方法,得到比较好的训练集。Shen Jie 等[8]在图形结构模型的基础上加上 CT(Clothing Technology)技术,即加入的衣服对图片的估计影响,加上 CT 的约束后可以提高估计的准确度。Brandon 等[9]提出了一种 And-Graph Model 来解决上述问题,是对图形结构的一种改进,并提出一

收稿日期:2016-02-10 **修回日期**:2016-06-15 **网络出版时间**:2016-11-21
基金项目:国家自然科学基金资助项目(61502005);安徽省自然科学基金(1308085QF100,1408085MF113)
作者简介:王国桢(1989-),男,硕士研究生,研究方向为图像处理和计算机视觉;方贤勇,教授,研究方向为计算机图形学和计算机视觉。

个基于边界的梯度特征(Histogram of Oriented Gradients,HOG)[10],然而对于一些复杂的背景图片,由于人不容易和背景图片区分出来,当检测图片中人的时候导致出现不准确的情况,从而影响最终的人体姿态估计结果。

针对已有的人体姿态估计对于复杂背景图片的不足,文中运用分割背景的思路,并与非常适合简单背景的 Sharp Context 的特征方法相结合,实现了高效的人体姿态的估计测量。

2 基于图模型的姿态估计

对于姿态的估计,文中要对图片进行训练和测试,因为对于复杂的背景,可能会影响对图像里面人的处理,从而影响估计结果。所以文中提出对图片中的人进行分割,去除背景的干扰,然后用形状上下文(Shape Context)[11]的特征检测方法进行训练和检测。

2.1 图像背景分割

文中使用文献[12]提出的一种基于卷积神经网络的分割方法对图像进行分割。该方法采用一种深度学习的方法训练模型和参数。而该方法可以把人从复杂的背景中分割出来,达到文中想要的结果。

首先要进行学习训练的过程,如图 1 所示,主要分成了特征提取和得到分类器的过程。特征提取是为了能获得合适的特征供分类器使用,从而可以在图像中把人分割出来。具体方法是:对于输入图像,首先用高斯滤波器并加入一个偏置量对输入图像进行卷积,得到卷积层。接着将它进行子采样,就是对卷积层中相邻四个像素求和使它变成一个像素,然后通过标量加权,再增加偏置,最后通过一个 sigmoid 激活函数,产生一个大概缩小四倍的特征映射图。整个过程可以看成是由卷积层和子采样层这两个层的交替组成。结束后就完成了对图像的特征提取。接着将得到的特征进行分类,获得一个分类器,这个分类器能够对输入图像进行初始的分割。然后,使用条件随机场的方法对上面得到的分类器效果进行提高,也就是相当于对一个粗糙的结果再进行优化的过程。通过这个过程,最终可以得到一个供文中使用的模型。接着输入图片,就可以得到文中想要的分割结果。

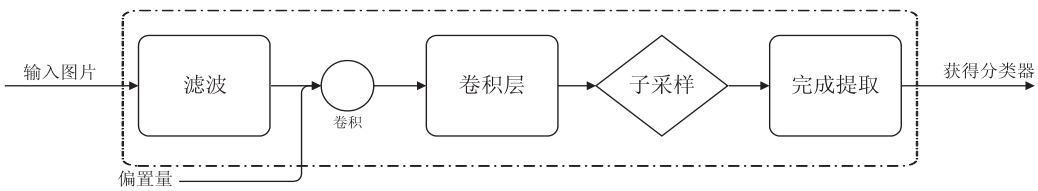


图 1 特征提取过程

对于得到的分割后的结果,可以用 Ehsan 等[13]提出的抠图方法。

对于图像中每个像素的颜色表示成前背景的线性组合。

$$H_z = a_z F_z + (1 - a_z) B_z \quad (1)$$

其中,F_z 是前景色;B_z 是背景色。

对于式(1)中的系数 a_z,它的取值范围是 0 ~ 1,称之为混合像素。一般抠图是需要用户手动交互信息的,但是文中有了之前求得的分割结果,能省去交互的过程,从而可以直接分割出结果,把人从背景中成功地抠出来,然后供后面训练等使用。

(a) (b)

图 2 图像分割及抠图结果

从图 2(a)中可以看出,从背景中把人给分割出来,然后由得到前景和后景的区别,可以用抠图的方法,把这个图中的人给抠出来,得到图(b),从而去除复杂的背景,方便之后的计算。

2.2 形状上下文特征提取及姿态估计

对于分割好背景的图,要对其进行边界提取,也就是使用形状上下文的方法,然后文中要用到图结构模型的方法来把人分块,分别对块内的特征进行寻找匹配,从而可以正确地估计出人体的姿态。

2.2.1 形状上下文

形状上下文可以很好地描述一个物体的形状特征,以测量形状的相似度。该方法主要是对轮廓上的 n 点,对于在 n 中的每一个点,用 p_i 和其他的 $n - 1$ 个点进行连接,从而可以获得 $n - 1$ 个向量。这一系列的向量对外形有着丰富的描述,可以直接决定形状的特征。所以如果当 n 的值很大时,所描述的特征也相对准确。

首先找到边缘上所有的点。文中可以用 Canny[14] 边缘检测算法获得边缘信息,取得二值图像,接着把所需要的轮廓给提取出来,这样就获得了图像中人的轮廓点。

把图像中所有点的坐标进行对数极坐标变换。对数坐标系建立后,把图像中的像素坐标从 (x,y) 转换成 (r,θ) ,然后要对极坐标系进行分割,将空间平均先分成 12 份,再以半径方向分成 5 份,这样空间就被分成 60 份,为 60 个单元 bin。以 p_i 为原点将整个图放到极坐标内,接着对每个 p_i 点求出它的直方图,也就是形状上下文,用 $M_i(k)$ 表示。其中,k 就是 bin 的序号,取值范围为 0 ~ 60,接着统计出每个 bin 中点的个数,然后绘制出一个直方图。最后就是对其余的点也做同样的操作,分别得到这些点的直方图,合到一起就得到了这个完整图像的形状上下文。

对于两个形状 W 和 U , p_i 是形状 W 上的任意一点, q_j 为形状 U 上的任意一点,则有 $C_s = C(p_i,q_j)$ 。其中, C_s 就是对两个形状上任意两点的匹配值,文中可以用 χ^2 检测(卡方检测),得到式(2):

$$C_s = \frac{1}{2}\sum_{k=1}^{K}\frac{[W(k) - U(k)]^2}{W(k) + U(k)} \tag{2}$$

文中希望求得式(2)的最小值,这样也就说明这两个点最为相似。有了形状上下文,就要计算两个形状的相似度。这里文中引用图结构模型。

2.2.2　图结构模型

图 3 为人体图模型。可以看出,主要就是把人体分成几个部块,有头、躯干、手臂和腿,其中手和腿分别又分成前手臂和后手臂,腿也是分成小腿和大腿。然后把这些部分块分别放入一个无向图中,使用图推理和概率学的知识对每一部分分别求解。

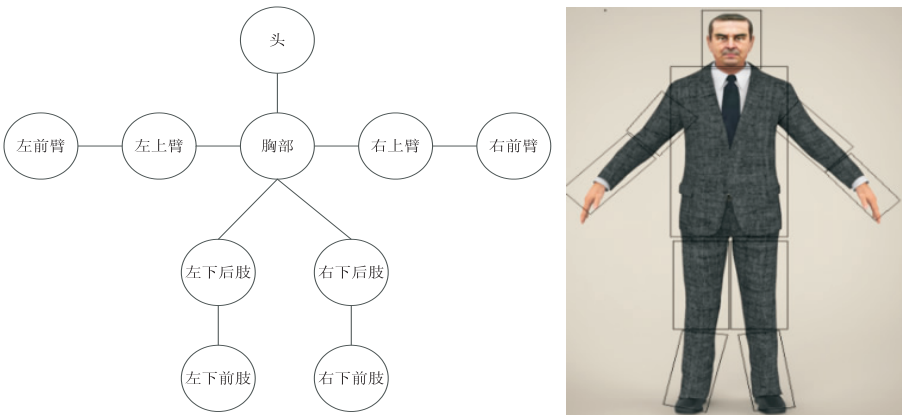


图 3　人体图模型

对于一个图结构模型,可以定义一个图 $G = (V, E)$,其中 $V = \{v_1,v_2,\cdots,v_n\}$ 表示各个顶点,相当于人体的各个部分。$(v_i,v_j) \in E$ 是连接 v_i 和 v_j 这两部分的一个边。每一个待检测的人,可以用 $H = (h_1,h_2,\cdots,h_n)$ 来表示,其中每一个 h_i 表示 v_i 的位置,也就是图 3 中黑色的矩形框, $h_i = (x_i,y_i)$ 表示矩形框中心位置的坐标。对于给定的一幅图, $m_i(h_i)$ 用来表示身体的真实部位和预测估计的 h_i 矩形框位置是不是符合的, $d_{ij}(h_i,h_j)$ 用来表示两个约束部位不会相差得太远。例如,头和身体就应该是相连接的而不可能是头和脚连接,这就需要进行约束。对于图像中的一个人,可以得出一个最小的能量优化式子:

$$X^* = \arg\min_X\left(\sum_{i=1}^{n} m_i(h_i) + \sum_{(v_i,v_j)\in E} d_{ij}(h_i,h_j)\right) \tag{3}$$

2.2.3　估计过程

文中要得到姿态估计的结果,首先要对图片进行大量训练。对于数据集中的每一张图片,文中都要先对它进行身体部位位置的标定,也就是框出身体的各个部位,然后求出它的形状上下文,即可以得到数据集中每一幅图像的形状上下文。

对于输入测试的图片,先对它求出形状上下文。由式(3)可知,文中要通过比较输入图像和数据集中的图像的形状上下文,使得这个式子得到最小值,就是所估计的结果。

对于式(3),文中可以用贝叶斯理论把它转化成概率学求解:

$$p(H/I) \propto p(I/H)p(H) \tag{4}$$

其中, I 为给定的一幅图像; $p(H)$ 为先验项。

$$p(H) = \prod_{(v_i,v_j)\in E} p(h_i \mid h_j) \tag{5}$$

它是为了确定人体的结构,对应的是公式(3)中的 $d_{ij}(h_i,h_j)$ 。

对于 $p(I/H)$ 有:

$$p(I/H) = \prod_{i=1}^{10} p(I/h_i) \tag{6}$$

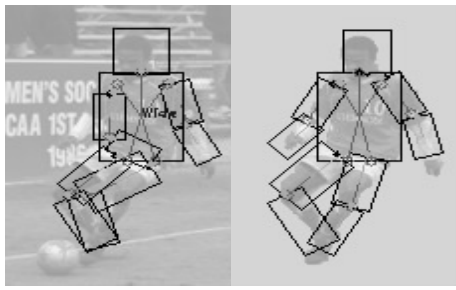
可以看出,该式就是为了求出人体的 10 个部位的位置,对应为式(3)中的 $m_i(h_i)$ 。当人体结构确定后,只需要通过式(2)中的对比形状上下文,对于输入图像找到最符合它的每一部分人体结构。把人体结构中的每一部分给填充上去,就可以得到人体的姿态估计。

3　实验和结果

为了验证上述方法的可行性和效果,分别对 Parse 数据集中 400 张图片进行训练和 100 张图片进行测

试,以及 Leeds 数据集中的 300 张图片分别进行训练和测试。实验主要以文献[7]的方法进行对比,因为文献[7]的方法也用到了图模型,然后在图模型的基础上用到 HOG 的特征来进行特征提取,并供其使用。文中方法和文献[7]方法的思路大体一样,所以选择和它进行对比。实验环境为:Window7 64 位操作系统,8 GB 内存,CPU 为酷睿 i7,软件为 Matlab 2015a。

图 4 是对静态图片中人体的姿态估计的结果。图(a)是文献[7]的方法,对于右手可以看出因为和背景的颜色过于相近,检测时可能就没法区分出袖子,所以没有取得很好的估计效果。图(b)是文中去除背景后加入形状上下文方法得出的效果,在右手的手部有了明显的提升,而且在左腿也比文献[7]方法的效果要准确。



(a)文献[7]的结果　　(b)文中结果

图 4　姿态估计的结果

图 5 和图 6 是文中方法和文献[7]方法在不同状态下的比较结果,分成人在跑步和静止状态下的图片对比。左图都是文献[7]的方法,右图都是文中方法。可以看出,文中的姿态估计的方法还是比文献[7]方法准确。

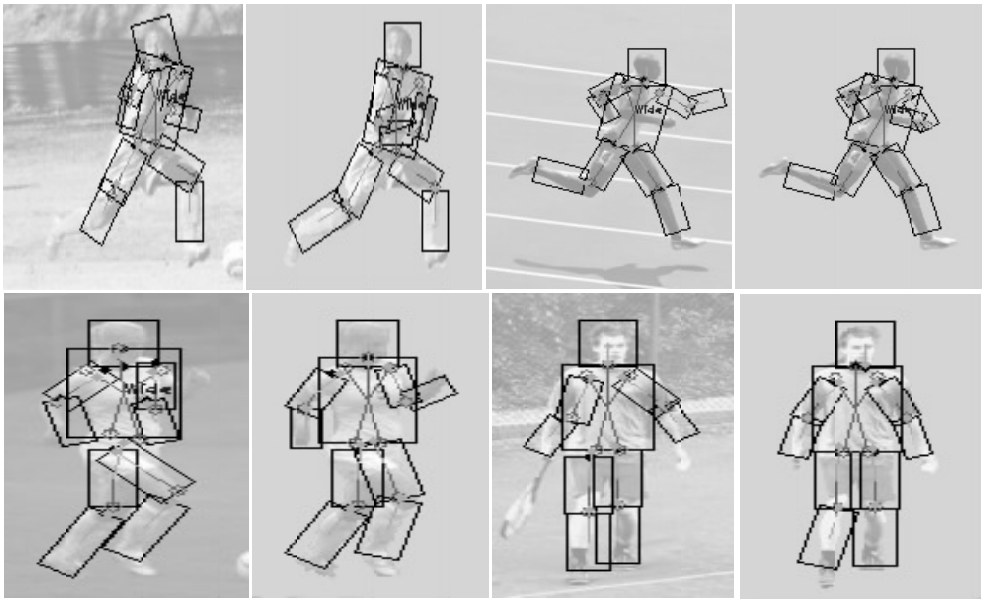


图 5　跑动的人的结果

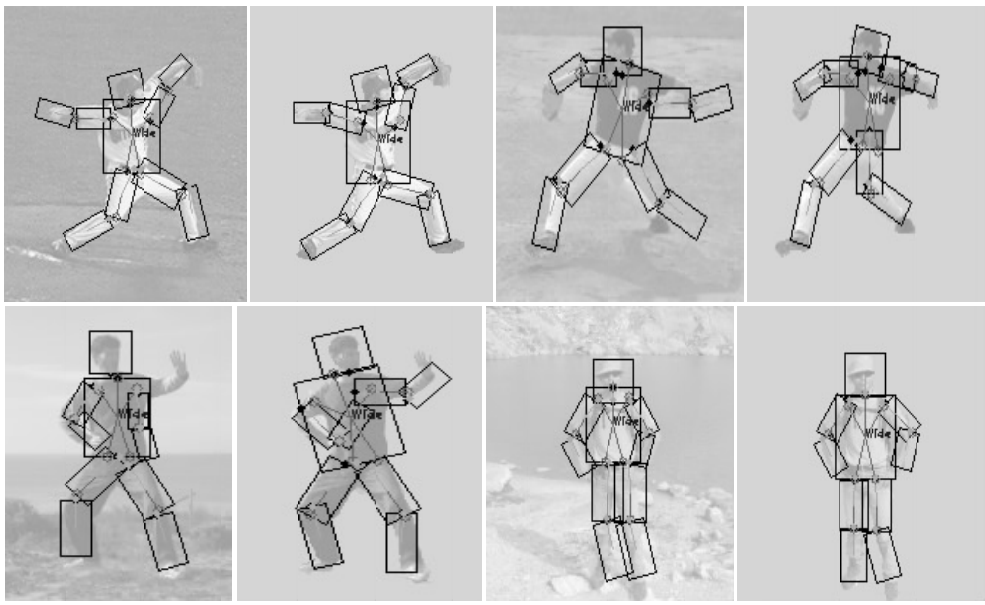


图 6　静止的人的结果

4　结束语

针对目前由于复杂背景不能很好地估计出人体姿态的问题,提出一种姿态估计方法。该方法首先去除图像的背景,然后再根据形状上下文的方法提取并训练样本模板,最后对输入图片进行对比,从而得出较好的姿态估计结果。实验结果表明该方法是可行的。但是可能或因为训练样本数不够,导致一些图像中的人没有得到很好的估计效果,以后工作中将会考虑加大训练集,并且对图模型结构进行改进,从而得到更好的姿态估计结果。

参考文献:

[1] Chairman-Hewett T T, Baecker R, Card S, et al. ACM SIGCHI curricula for human-computer interaction[R]. New York: ACM, 1992.

[2] Myers B A. A brief history of human-computer interaction technology[J]. Interactions, 1998, 5(2): 44-54.

[3] Felzenszwalb P, Huttenlocher D. Pictorial structures for object recognition[J]. International Journal of Computer Vision, 2005, 61(1): 55-79.

[4] Yao B Z, Nie B X, Liu Zicheng, et al. Animated pose templates for modelling and detecting human actions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 436-452.

[5] Chaudhry R, Ravichandran A, Hager G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions[C]// IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2009: 1932-1939.

[6] Ouyang Wanli, Chu Xiao, Wang Xiaogang. Multi-source deep learning for human pose estimation[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2014: 2337-2344.

[7] Ge Hinton, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 21(18): 1527-1554.

[8] Shen Jie, Liu Guangcan, Chen Jia, et al. Unified structured learning for simultaneous human pose estimation and garment attribute classification[J]. IEEE Transactions on Image Processing, 2014, 23(11): 4786-4798.

[9] Rothrock B, Park S, Zhu Songchun. Integrating grammar and segmentation for human pose estimation[J]. International Journal of Computer Vision, 2013, 25(13): 3214-3221.

[10] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2005: 886-893.

[11] Belongie S, Malik J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis and Machine intelligence, 2000, 18(5): 927-944.

[12] Zheng Shuai, Jayasumama S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2015: 216-232.

[13] Shahrian E, Rajan D, Price B, et al. Improving image matting using comprehensive sampling sets[C]//IEEE conference on computer vision and pattern recognition. [s. l.]: IEEE, 2013: 636-643.

[14] Canny J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligene, 1986, 8(6): 679-698.