深度图像中基于部位位置及尺寸的人体识别方法

赵文闯 程 俊

（ 中国科学院深圳先进技术研究院 深圳 518055 ）

摘 要 人体姿态估计方法中，在初始化或者跟踪失败的情况下，需要提供姿态初始值。我们将姿态估计看作对人体每个像素的分类问题，设计了一种表征人体部位尺寸及位置的特征。通过识别当前帧人体像素所属部位，可计算人体姿态。我们对分类器性能进行了测试，分类器对人体像素的识别率达到91%，对分辨率为160*120的深度图像，Intel单核1.6 GHZ的处理器上的处理速度为4 ms/fps。本文分析了该特征的局限性及出现问题的原因。

关键词 人体姿态估计；随机森林

Human Body Recognition from Depth Image Based on Part Size and Position

ZHAO Wen-chuang CHENG Jun

(*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)

Abstract Human action recognition acts as an important role in human machine interaction. This paper proposes a human body recognition method from depth image based on part size and position features. Random forest classifiers are trained with different parameters. Experimental results demonstrate the feasibility of proposed approach. Recognition accuracy is about 91% and the computation time is about 0.96 us per pixel.

Keywords human body recognition; random forest

1 引 言

人体姿态估计是当前计算机视觉领域的一个研究热点，在深度图像中估计人体姿态，首先定义模型和观测之间的误差测度，搜索姿态参数使得人体模型与观测之间的误差最小[1]。

由于误差函数具有多个局部极值[1]，在误差最小化过程，粒子群[2,3]、层次粒子群[4]、层次进化[5]等智能优化算法用于求解全局最优，但是算法的实时性较差。为提高速度，Mussi等[3]还在GPU上实现了粒子群算法。考虑到算法的实时性，不可能在整个姿态空间中搜索求解最优姿态，一般采用上一帧的姿态作为初值，在初值局部进行搜索[1]，常用的方法主要有基于贝叶斯框架的方法[6-10]，以及基于关节体配准的方法[11-14]，Siddiqui[10]指出，基于贝叶斯框架的方法效果略微好些，但速度较慢，基于关节体配准的方法速度比基于贝叶斯框架的方法速度快接近90倍。局部搜索进行姿态估计的方法，运行速度较快，也能达到较高的精度，但需要手动初始化，并且当人体运动剧烈时，该方法容易陷入局部最优。PrimeSense要求用户作出“投降”姿势完成初始化，为处理局部最优的问题，Andriluka等[15]使用了检测跟踪法（Tracking by Detection）。

微软剑桥研究院Shooton[16]将人体姿态估计看做每个像素的分类问题，即计算每个像素属于人体模型中每个部位的概率，对像素进行组合，得到每个部位的像素，再生成关节位置的假设。该方法获得了不错的效果，并且可为局部搜索方法提供初始值。然而，该方法计算速度较慢，在8核处理器上处理速度为5 ms/fps，并且该方法需要的样本数量巨大，据微软介绍样本量为TB级。

基金项目：本项研究由中国科学院院地合作项目（ZNGZ-2011-012），深圳市深港创新圈项目（JSE201007200037A）资助。

作者简介：赵文闯，工程师，主要研究方向为模式识别，E-mail：wch.zhao@siat.ac.cn；程俊，研究员，博士生导师，主要研究方向为计算机视觉、模式识别。

本文通过对每个像素分类来估计人体姿态，基于各个部位在人体内部位置分布和尺寸大小的差异，设计了一种表征人体部位尺寸及位置的特征，通过使用适量样本进行学习，识别人体表面每个像素的类别。本文内容安排如下：第二节描述了本文使用的表征部位位置和尺寸差异的特征；第三节中，我们基于本文特征训练随机森林分类器，测试了分类器的性能并分析了本文特征的局限性；第四节对本文内容进行了总结。

2　特征描述

本文采用boardcard人体模型，分为躯干、头、左上臂、左下臂、右上臂、右下臂、左大腿、左小腿、右大腿、右小腿等十个部位，如图1所示。

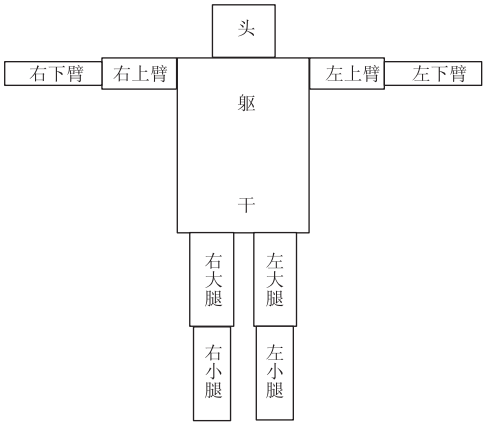


图1　人体模型示意图

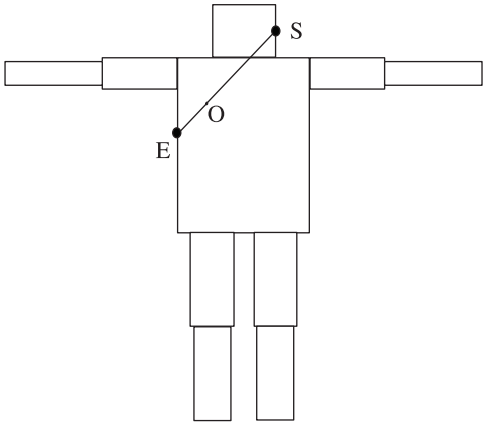


图2　扫描线示意图

各个部位的位置分布不同，例如多数情况下头部在躯干上方，而左上腿、左下腿、右上腿、右下腿分布在躯干下方。人体部位尺寸也不同，例如躯干尺寸要远大于左、右下臂。

基于以上特点，本文设计了表征人体部位位置及尺寸的特征。

（1）人体部位位置特征

直接使用像素的三维坐标描述位置特征，设像素坐标为(x,y,z)，表征像素位置的特征向量为$[x,y,z]^T$。

（2）人体部位尺寸特征

对于人体深度图像中的像素，本文使用经过该像素的扫描线段的长度作为表征该像素所在部位的尺寸。为避免插值，使用0度，45度，90度，135度，180度，225度，270度，315度等八个方向的扫描线段。

由于m度方向和m+180度方向构成一条直线，扫描线段定义为从m度方向的第一个跳变开始到m+180度第一个跳变结束的全部像素，这里的跳变是指扫描方向上两个相邻像素之间，从背景像素变化为人体像素或者从人体像素变化为背景像素。如图2所示，过O点45度方向的第一个跳变像素为S，225度方向的第一个跳变像素为E，则过O的一个扫描段为SE。

扫描段长度定义为扫描段两个断点的欧式距离，设起始像素三维坐标(x_s,y_s,z_s)，结束像素三维坐标(x_e,y_e,z_e)，则扫描的尺寸为：

$$d=\sqrt{(x_s-x_e)^2+(y_s-y_e)^2+(z_s-z_e)^2} \quad (1)$$

对每个像素，本文采用4个扫描线段的长度表征像素所在部位尺寸的大小，则表征部位尺寸的特征向量为$[d_0,d_{45},d_{90},d_{135}]^T$，其中，$d_0$表示过该像素的水平方向扫描线段的尺寸，$d_{45}$表示过该像素的45度方向扫描线段的尺寸，$d_{90}$表示过该像素的垂直方向扫描线段的尺寸，$d_{135}$表示过该像素的135度方向扫描线段的尺寸。

（3）特征归一化

对每个像素，得到特征向量$[d_0,d_{45},d_{90},d_{135},x,y,z]^T$，需要对特征向量进行尺寸归一化和位置归一化。

位置归一化过程，先进行PCA分析，再进行PCA重投影，对投影后坐标除以身高$height$。尺寸归一化采用扫描线段的长度除以身高$height$，即$[d_0/height,d_{45}/height,d_{90}/height,d_{135}/height]^T$。

这里，身高通过PCA过程中协方差矩阵的最大特征值估计得到，我们认为，身高正比于最大特征值的平方根，即

$$height=k\times\sqrt{\lambda_{\max}} \quad (2)$$

采集样本进行回归分析，得到k=4。

3　实验与分析

本节中，我们采用基于关节体配准算法进行人体像素部位标记，采用三个测试集测试分类器性能，并对分类器性能进行了分析。此外，我们还分析了本文特征的局限性及原因。

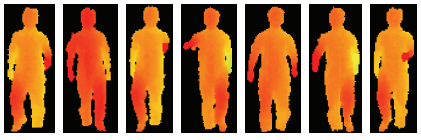
（1）样本标记

本文采用PrimeSense深度图像传感器采集了人体常见的运动过程，使用基于关节体配准算法标记求解人体姿态，对于人体点云中的像素，设定该像素所属部位为距离该像素最近的人体部位，因此，样本标记相对粗糙，对于关节处的像素尤其如此。为避免标记不准确的问题，在训练时我们使用每个部位的非边界像素作为样本。

（2）测试集描述

我们设计了三个测试样本集，样本集1中人躯干直立，仅手臂运动，共约600个测试图像，部分测试样本图像如图3（a）所示；样本集2中人做走路的动作，共约500个测试图像，部分测试样本图像如图3（b）所示；样本集3中，人自由运动，有点接近于跳舞，共1000个测试图像，部分测试样本图像如图3（c）所示。由于我们不处理遮挡的情况，故三个样本集基本上无遮挡严重的图像。

（a）测试集1，躯干不动，手臂运动



（b）测试集2，走路

（c）测试集3，自由运动，接近于跳舞

图3 样本测试集

（3）分类器性能

随机森林分类器是一个多类分类器，具有学习、识别速度快，不易过拟合的优点，因此，本文采用随机森林分类器。

在随机森林中，主要的参数包括森林中树的个数以及每颗树的深度，使用不同数量的树（3~9棵）和不同的深度（6~12层）训练分类器并使用前述三个测试集测试分类器性能。由于识别速度正比于树的个数以及树的深度，本文重点关注识别率指标。不同参数下，分类器对三个测试集得分辨率如图4所示。

树的深度对识别率起至关重要的作用，树的个数可防止训练过程过拟合。从图4中可以发现，当树的深度达到9层或以上时，对3~9之间的任意树的数量，分类器对三个样本集的识别率均达90%以上，并且识别率随着树的深度增加而提高，这种现象在测试集3中表现尤为明显。

使用测试集3对分类器进行测试时，深度为k+1的

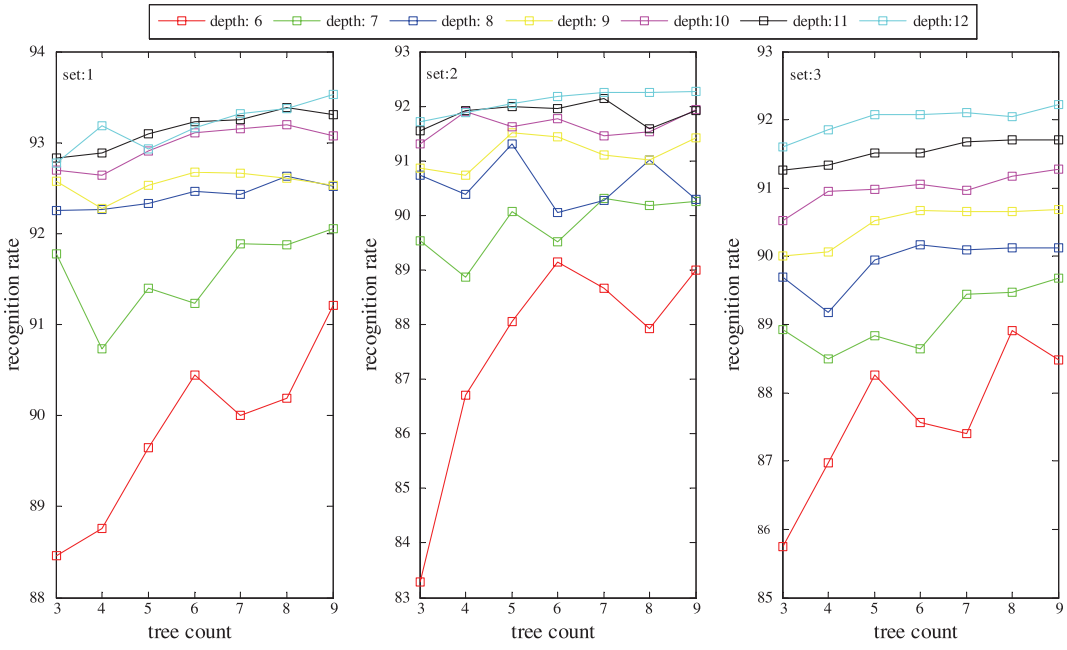


图4 不同参数下分类器识别率

7个分类器识别率最小的分类器的识别率要高于或接近深度为k的分类器的7个分类器的最大识别率，这是因为测试集3包括的各种人体姿势，这就要求分类器本身比较复杂，深度更大。

设定树的深度为10，比较相同层数、不同树的个数的分类器对三个测试集的识别率，对测试集1，识别率在92.5%到93.0%之间，对测试集2，识别率在91%到92%之间，对测试集3，识别率接近91%。可以发现，三个测试集中，识别难度逐渐增加，这是因为：

①测试集1躯干和腿静止，仅手臂运动对人体中各个像素的位置分布影响较小。

②测试集2中，手臂和腿均运动，对人体中各个像素的位置分布影响相对较大，并且人走路过程手臂和躯干相隔较近，并且会偶尔遮挡躯干，导致识别难度增加。

③测试集3中，人体运动非常复杂，并且会出现一些不常见的姿势，对人体中各个像素的位置分布影响更大，识别难度最大。

综合考虑识别率和识别速度，采用4棵树，10层随机森林分类器，对三个测试集得识别率分别为：92.64%，91.90%，90.95%。

在Intel单核1.6 GHZ的处理器上测试识别速度，不考虑特征提取时间，对于160×120的深度图像，平均特征提取及识别速度为1.86 us/pix，人体距深度传感器2.5 m时，身高为1.7 m的人体上大约有2000个点，处理时间约为4 ms。

本特征为7维向量，重要性依次为34.4%，24.0%，7.4%，6.8%，9.2%，9.6%，8.5%，位置的重要性接近70%，识别错误发生主要发生在相邻部位之间。以测试集3为例，x坐标为样本真实类别，y坐标分类器对样本的识别类别，z坐标为识别错误的比例，说明识别错误分布，如图5所示。图5中，图中每种颜色表示一个真实的部位，每个真实部位和每个识别部位对应一个立方体，该立方体表明识别错误占全部识别错误的比例。

从图5可以看出，头被识别为躯干（7.67%），左上腿被识别为躯干（6.85%），右上腿被识别为躯干（6.04%），左下腿被识别为左上腿（6.89%），左上臂被识别为躯干（5.77%），右上臂被识别为躯干（6.15%）等，这是相邻部位的活动区域接近，尺寸

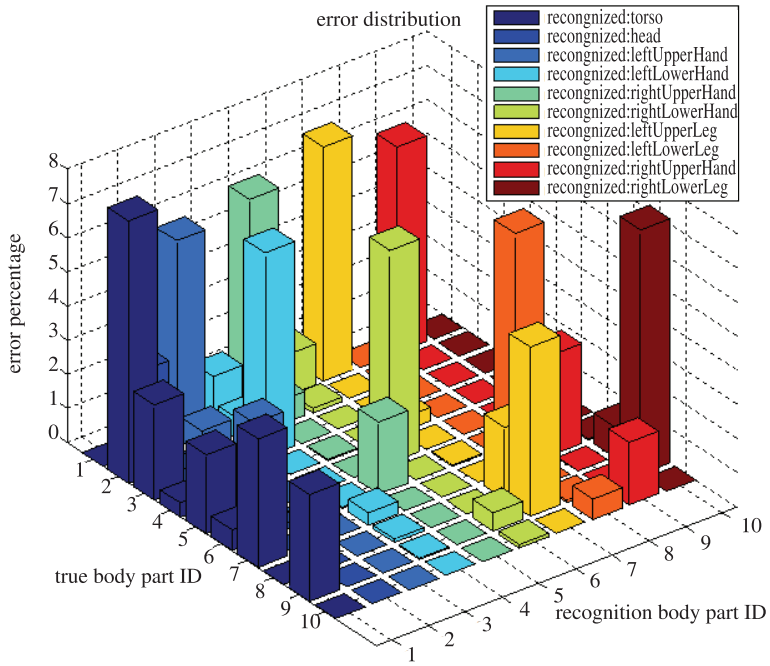


图5　识别错误分布图

也差别不是很大的缘故。

（4）局限性

由于位置的重要性接近70%，当人体变形较大时，位置归一化效果会差，如图6所示，若左腿上像素位于中轴的右侧偏远，会被识别为右腿，若小腿略微抬起时，小腿上的部分像素会被识别为大腿。

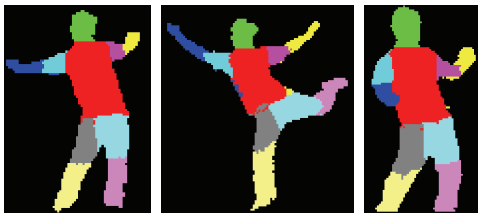


图6　部分识别出错的情况

出现这种问题的原因是归一化过程仅使用人体全局信息，在特征向量中增加像素的局部信息将有助于缓解该问题。

本文关注非遮挡情况下的人体识别，遮挡问题处理可在后处理过程中完成。实际上，如果在训练集中加入遮挡样本，由于位置的重要性接近70%，该特征可以处理一部分的遮挡情况。但对于部位A遮挡部位B，并且A、B距离较近的情况，部位A的像素会识别为部位B的像素，针对此情况，需要引入图像序列信息。

4 结 论

基于人体部位的位置及尺寸上的差异，本文设计了一种表征人体部位尺寸及位置的特征，使用随机森林分类器进行学习，经测试，分类器对人体像素的识别率达到91%，对分辨率为160×120的深度图像，Intel单核1.6 GHZ的处理器上的处理速度为4 ms/fps。本文还分析了所设计特征的局限性以及导致该问题的原因。

参 考 文 献

[1] Poppe R. Vision-based human motion analysis: An overview [J]. Computer Vision and Image Understanding, 2007, 108(1-2): 4-18.

[2] Ivekovic S, Trucco E, Petillot Y R. Human body pose estimation with particle swarm optimization [J]. Evolutionary Computation, 2008, 16(4): 509-528.

[3] Mussi L, Ivekovic S, Cagnoni S. Markerless articulated human body tracking from multi-view video with GPU-PSO [C] // Proceedings of the 9th International Conference on Evolvable Systems: from Biology to Hardware, 2010.

[4] Vijay J, Ivekovic S, Trucco E. Articulated human motion tracking with HPSO [C] // Proceedings of The Fourth International Conference on Computer Vision Theory and Applications, 2009: 531-538.

[5] Robertson C, Trucco E. Human body posture via hierarchical evolutionary optimization [J]. Image and Vision Computing, 2010, 28(11): 1530-1547.

[6] Sidenbladh H, Black M J, Fleet D J. Stochastic tracking of 3d human figures using 2d image motion [C] // Proceedings of the 6th European Conference on Computer Vision-Part II, 2000: 702–718.

[7] Sminchisescu C, Triggs B. Covariance scaled sampling for monocular 3d body tracking [C] // Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA: IEEE Computer Society Press, 2001.

[8] Pavlovic V, Rehg J M, Cham T J, et al. A dynamic bayesian network approach to figure tracking using learned dynamic models [C] // Proceedings of IEEE International Conference on Computer Vision, 1999: 94–101.

[9] Zhu Y D, Fujimura K. A bayesian framework for human body pose tracking from depth image sequences [J]. Sensors, 2010.

[10] Matheen Siddiqui, Gerard Medioni. Human pose estimation from a single view point, real-time range sensor[C] // Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[11] Moschini D, Fusiello A. Tracking stick figures with hierarchical articulated ICP [C] // Proceedings of the First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences, 2008: 61–68.

[12] Ni B B, Winkler S, Kassim A. Articulated object registration using simulated physical force/moment for 3D human motion tracking [C] // Proceedings of the 2nd Conference on Human Motion: Understanding, Modeling, Capture and Animation, 2007: 212-224.

[13] Demirdjian D, Darrell T. 3-D articulated pose tracking for untethered diectic reference [C] // Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, 2002: 267–272.

[14] Demirdjian D. Enforcing constraints for human body tracking [C] // Conference on Computer Visionand Pattern Recognition Workshop, Madison, Wisconsin, USA, 2003, 9: 102–109.

[15] Andriluka M, Roth S, Schiele B. Monocular 3D pose estimation and tracking by detection [C] // Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[16] Shotton J, Fitzgibbon A, Moore R, et al. Real-time human pose recognition in parts from single depth images [C] // Computer Vision and Pattern Recognition Workshops (CVPRW), 2011.