基于深度图像的多学习者姿态识别

张鸿宇 刘 威 许 炜 王 辉
(华中科技大学电子与信息工程系 武汉 430074)

摘 要 在数字化学习场景中,人体姿态的识别有助于分析学习者的学习状态。提出了一种基于深度图像的多学习者姿态识别方法。首先通过Kinect的红外传感器获取包含深度信息的图像,利用深度图像进行人像-背景分离;然后提取人体的轮廓特征Hu矩;最后采用SVM分类器对轮廓特征进行分类和识别。实验结果表明,本方法能有效地识别多个学习者的举手、正坐和低头等姿态。

关键词 姿态识别,深度图像,多学习者
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.9.059

Depth Image Based Gesture Recognition for Multiple Learners

ZHANG Hong-yu LIU Wei XU Wei WANG Hui
(Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract Gesture recognition of the learner's body is helpful for analysing and evaluating the learner's status in the e-learning system. In this paper, a depth image based gesture recognition method was proposed to recognize multiple learners. After obtaining the depth image from Kinect sensor, the human body is separated from the background image and the contour features described in Hu moments are extracted. The learner's gesture is then recognized based on SVM classifier. The test results show that this method can efficiently recognize the hand-up, sitting and head-yield gestures of multiple learners.

Keywords Gesture recognition, Depth image, Multiple learner

1 概述

随着远程教育的应用发展,如何有效地记录和评估数字化学习过程中学习者的状态成为日益重要的课题。目前已经有多种检测学习者状态的系统方案,例如,利用面部表情识别获取用户心理状态[1],利用眼睛状态判断用户疲劳程度[2]等。这些系统通过智能检测学习者的学习状态,让教师得到更多的反馈信息,从而达到改进教学和提高学生学习效率的目的。值得注意的是,除了学习者的面部表情,身体姿态也部分体现了学习者的学习状态。例如,当学生学习状态积极时坐姿就端正,老师提问时就会举手发言;当学生学习状态消极时就会打瞌睡、趴在桌子上等。学习者的姿态信息的测量与记录,对于教师事后分析学习者的学习状态、改进教学过程具有重要作用。

人体姿态识别技术通过对传感器采集的数据进行加工、处理和分析,使得计算机系统能够理解个体动作、个体之间以及个体与环境之间的交互[3]。人体姿态识别的传统方法是通过2D图像进行人体前景提取,然后基于人体模型进行姿态估计。如徐光祐等[4]提出利用骨骼和关节模型建立人体上肢模型,通过对18个未知参数进行估计,得到人体姿势;Triggs等[5]提出基于样本的方法,使用非线性回归从单目图像轮廓中直接估计3D姿势等。然而,这些单目视觉的方法存在一定的局限性。首先,图像的深度信息被丢失,在图像理解方面容易产生歧义;其次,此类方法对光照强度较为敏感。

随着体感测量技术的日益成熟,部分装置(例如微软推出的Kinect产品)可以通过红外传感装置测距并提供部分骨骼检测的信息。这些信息的引入能够显著提高现有人体姿态识别算法的识别效果。例如Hariharan等[6]在教学场景中基于人体关节点信息来识别学生手势;Hsu等[7]在远程会议场景中基于人体关节点信息对与会人员进行手势识别;Vermun等[8]研发了基于姿态识别结果的教学反馈系统。需要指出的是,这些工作依赖于人体关节信息的获取,而目前Kinect装置最多仅能提供两人的骨骼信息的检测,因此这些方案最多只能识别两个学习者的姿态信息。

本文针对数字化学习场景中多学习者姿态识别的问题,提出了基于深度图像的检测方法,即直接利用Kinect的深度信息,对图像进行人像-背景分离,结合人脸检测技术对人体轮廓进行筛选校正,然后提取轮廓特征Hu矩,通过SVM对轮廓特征分类,实现了一种多学习者情况下的人体姿态估计方法。本方法克服了现有工作[7,8]只能识别最多两人姿态的局限性,实测结果表明本方法可以支持4个学习者的姿态识别。

到稿日期:2014-09-22 返修日期:2015-01-12 本文受国家科技支撑计划项目(2013BAH72B01-1)资助。
张鸿宇(1990—),硕士生,主要研究方向为视觉处理;刘 威(1977—),博士,副教授,主要研究方向为视觉处理、互联网应用等;许 炜(1977—),博士,副教授,主要研究方向为互联网应用;王 辉(1991—),硕士生,主要研究方向为视觉处理。

本文第2节介绍了整个系统的设计思路；第3节详细介绍了系统的关键模块实现；第4节是测试结果和分析；最后总结全文。

2 系统设计

2.1 技术路线

本文设计的系统框架如图1所示，共包括3个主要模块，分别是数据采集模块、人体轮廓提取模块和特征提取和判别模块。

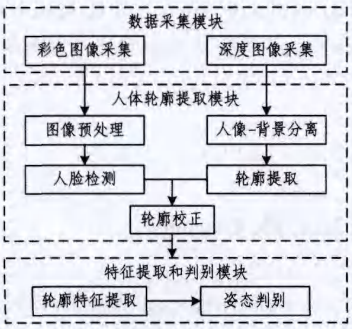


图1 系统流程

首先，数据采集模块用Kinect同时采集彩色图像和深度图像。接下来，人体轮廓提取模块从深度图像中提取具有人体形状的轮廓，同时从彩色图像中提取人脸，通过两类信息联合判别去除非人体轮廓。最后，特征提取和判别模块提取轮廓Hu矩特征，并利用已经训练好的SVM分类器对不同轮廓进行分类和识别。

2.2 软硬件环境

本系统的软件开发平台是Visual Studio 2010，开发语言是C++，主要图像处理算法基于第三方计算机视觉库OpenCV实现。

本系统的硬件测量部分采用了Kinect传感器。该装置是微软开发并应用于xbox游戏机主机的体感设备。Kinect基本组成及功能如表1所列。

表1 Kinect结构和功能

硬件结构	功能介绍
RGB摄像头	获取彩色图像，自身视角垂直方向43°，水平57°
红外发射装置	发射红外线，形成散斑图像，测量距离0.8m～3.5m
红外摄像机	接收散斑图像信息，并创建深度图像
阵列麦克风	接收声音信息

Kinect传感器是一个外形类似网络摄影机的装置。Kinect有3个镜头：中间的镜头是RGB彩色摄影机，左、右两边镜头则分别为由红外线发射器和红外线CMOS摄影机所构成的3D结构光深度感应器。Kinect红外测距部分的有效采集距离是0.8m～3.5m。

3 关键模块

3.1 数据采集模块

本系统用到了Kinect的带用户编号的深度信息，深度数据结构如图2所示。Kinect采集到的每个像素的深度数据用两个字节来保存，高13位保存了深度值，低3位保存用户序号，通过提取用户序号就能很好地将人体和周围的环境分离开来。

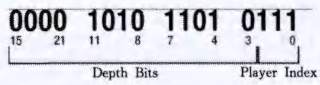
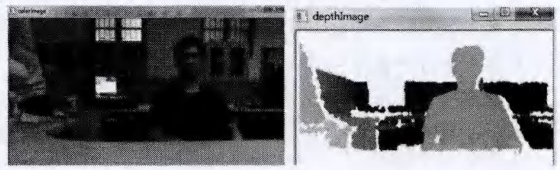


图2 深度数据结构

Kinect采集到的图像数据如图3所示，灰度范围为0—255。其中图3(a)是原始的彩色图像，图3(b)是对应的原始深度图像。从图中可以看到，在Kinect有效距离(0.8m～3.5m)内的物体深度信息能捕捉到，超出有效距离的像素值都被设置成了255，即远端背景呈现的白色。



(a)彩色图像 (b)深度图像

图3 Kinect采集的彩色图和对应的深度图

3.2 人体轮廓提取模块

3.2.1 人像-背景分离

通过Kinect得到场景的深度图像之后，利用这一图像信息进行人像-背景分离。传统的人像-背景分离通常是基于高斯混合模型的背景建模方法，这种方法的弱点在于：当人体处于背景变化过快或者是光照条件变化较明显的环境中时，分离的效果较差。

为了解决这个问题，本文利用Kinect提供的深度信息实现人像-背景的分离。具体做法是：对于每一个像素，根据其深度数据去修改其图像的RGB值。如果属于同一个用户的ID，那么像素就标为同种颜色；不同的用户，其ID不一样，颜色的标识也不一样；如果不属于某个用户的像素，那么就采用原来的深度值。

图4给出了一组处理结果的示意图。其中图4(a)是采集的原始彩色图像；图4(b)是结合Kinect标识的用户ID进行区分后得到的彩色深度图像；图4(c)是根据彩色标识值进行区分后得到的人像和背景分离后的图像。可以看到，本文方法将人体和背景进行了有效的分离。

(a)彩色图像

(b)标记人体后的深度图

(c)人像和背景分离后的图像

图4 人体轮廓提取示意图

3.2.2 轮廓提取和处理

通过不同颜色对人体和背景进行标识后，进一步提取人体轮廓。其基本原理是将图像转化为二值化图像，通过图像形态学处理去除杂点，然后通过二值图像提取算法提取轮廓。

具体步骤是：首先对图像中的所有轮廓进行检测。从图像左上角开始遍历图像，若某像素周围 8 个点的像素值全为 0 或者全为 255，则判断该点为内部点；若某像素周围 8 个点的像素值既有 255 也有 0，则判断该像素为轮廓点；判断获得的内部点的像素值置为 0，轮廓点的像素值置为 1。然后，根据系统预设的检测门限判定待检测人体的轮廓面积阈值，去除掉相较于阈值过小的轮廓，最终得到清晰完整的人体轮廓。

图 5 给出了图 4 对应图像的轮廓提取结果。可以看出原图 4(c)中右起 3 个用户的身体形态内外的杂点轮廓已经被剔除，在图 5 中获得了较为完整的人体轮廓。

图 5 人体轮廓提取结果

3.3 特征提取和判别模块

3.3.1 基于 Hu 矩的轮廓特征的提取

不变矩是一种高度浓缩的图像特征，在目标检测与识别中有着广泛的应用。1961 年，Hu 提出由 7 个归一化中心矩组合成的矩[9]，即 Hu 矩。Hu 矩可以较好地表征轮廓的外形，并且具有平移、光照和旋转等多畸变不变性[10]。因此，本系统采用 7 个不变矩构成的特征量作为人体轮廓的特征向量。

假定连续情况下图像函数为 $f(x,y)$，则该图像的 $p+q$ 阶几何矩(即标准矩)m_{pq} 的定义为：

$$m_{pq}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^p y^q f(x,y)\mathrm{d}x\mathrm{d}y,\ p,q=0,1,2,\cdots \tag{1}$$

该图像的 $p+q$ 阶中心距 μ_{pq} 的定义为：

$$\mu_{pq}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\bar{x})^p(y-\bar{y})^q f(x,y)\mathrm{d}x\mathrm{d}y,\ p,q=0,1,2,\cdots \tag{2}$$

对于离散的数字图像，采用求和号代替积分，以 $(\bar{x},\bar{y})$ 代表图像的重心，以 N 和 M 分别标记图像的高度和宽度，则相应的标准矩和中心矩为：

$$m_{pq}=\sum_{y=1}^{N}\sum_{x=1}^{M}x^p y^q f(x,y),\ p,q=0,1,2,\cdots \tag{3}$$

$$\mu_{pq}=\sum_{y=1}^{N}\sum_{x=1}^{M}(x-\bar{x})^p(y-\bar{y})^q f(x,y),\ p,q=0,1,2,\cdots \tag{4}$$

在离散图像的标准矩和中心矩基础上，可以定义归一化的中心距：

$$\eta_{pq}=\mu_{pq}/(\mu_{00}^{\rho}) \tag{5}$$

其中，$\rho=(p+q)/2+1$。

Hu 矩的 7 个归一化中心矩就是在二阶和三阶归一化中心矩的基础上构造出来的，在此标记为不变矩 M_1-M_7：

$$M_1=\eta_{20}+\eta_{02} \tag{6}$$

$$M_2=(\eta_{20}-\eta_{02})^2+4\eta_{11}^2 \tag{7}$$

$$M_3=(\eta_{30}-3\eta_{12})^2+3(\eta_{21}-\eta_{03})^2 \tag{8}$$

$$M_4=(\eta_{30}+\eta_{12})^2+(\eta_{21}+\eta_{03})^2 \tag{9}$$

$$M_5=(\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2+(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2] \tag{10}$$

$$M_6=(\eta_{20}-\eta_{02})[(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]+4\eta_{11}(\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03}) \tag{11}$$

$$M_7=(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]-(\eta_{30}-3\eta_{12})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2] \tag{12}$$

本文在系统中实现了对轮廓图像的分离，然后对每个轮廓分别计算其 Hu 矩特征值，这些计算结果将作为后续对人体姿态判别的数据基础。

3.3.2 基于 SVM 的轮廓特征分类

在教室环境下人体的坐姿往往会有较多的遮挡，这给完整的人体姿态估计带来了困难。考虑到系统目标是区分学习者在学习活动中的活动表现，只需分辨出在学习活动中常见的坐姿。通过分析，判定了 3 种不同的坐姿(见图 6)，通过对学生坐姿的分析发现，学生的这 3 种坐姿在学生的人体轮廓上有比较明显的区别度。可以通过学生的轮廓对这 3 种坐姿(即举手、正坐和低头)进行识别。

图 6 3 种姿态的轮廓样本

我们选取图 6 所示的姿势作为本系统人体姿态的基准，计算得到的 3 种坐姿轮廓的 Hu 不变矩如表 1 所列。

表 1 3 种姿态人体轮廓样本的 Hu 矩

姿态	Hu1	Hu2	Hu3	Hu4	Hu5	Hu6	Hu7
正坐	2.04 e−001	6.87 e−003	3.61 e−003	2.63 e−004	1.89 e−007	1.41 e−005	1.73 e−007
举手	1.99 e−001	2.76 e−003	4.10 e−003	1.94 e−004	1.73 e−007	1.02 e−005	5.34 e−009
低头	1.90 e−001	8.31 e−003	8.75 e−004	4.77 e−005	4.61 e−009	2.41 e−006	8.59 e−009

考虑 Hu 矩是一组轮廓的向量特征，可以选用一些已有的分类算法，将 Hu 矩作为训练数据来训练分类器，然后用测试数据对姿态估计的结果进行评估。在实验中，本文采用了支持向量机 SVM 作为分类器，目前实验中是一个三分类的问题，鉴于将来可能扩展到多种姿态类别，本文选用的是一对一 SVM 分类方法。

该方法首先由 Knerr 提出，也叫 Pairwise Method[11]。其基本思路是在 n 类问题中进行两两组合，构造 $n(n-1)$ 个二分类器，每个分类器只针对两类数据进行训练。分类时，将待分类样本分别输入 $n(n-1)/2$ 个二类分类器，各个分类器分别对其识别，最后这些分类结果对各个类别进行投票，得票最多的类别为最终的识别结果。这种方法的优点是训练速度快，在类别数目不大时效果较好；缺点是当类别数目过大时产生的子分类器过多，性能下降明显。对于本系统而言，该分类方法是适用的。

4 结果测试

4.1 场景设置

所搭建的多学习者姿态识别的实验场景如图 7 所示。其中 Kinect 采集装置放置在教师讲台附近的位置，保持正对待检测的学习者。

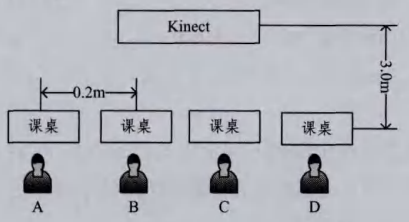


图 7 实验场景示意图

在实验环境中，学生与 Kinect 的距离为 3.0m，座位之间相隔 0.2m。这样既能保证待检测学生在 Kinect 视野范围内，也保证了人体姿态的区分度。为了让实验环境更加接近真实的远程教室场景，座椅前后错开放置。

4.2 实验结果

从左至右将学生依次编号为 A、B、C、D；其姿态举手、正坐、低头依次标记为状态 1、2、3，未知姿态标记为状态 0。例如，若 A 学生举手，则程序判别为“A-1”；若 A 学生正坐，则程序判别为“A-2”；若 A 学生低头趴着，则程序判别为“A-3”；若无法判别，则输出“A-0”。

本文开展了多组实验，其中的 3 组实验结果如图 8 所示。

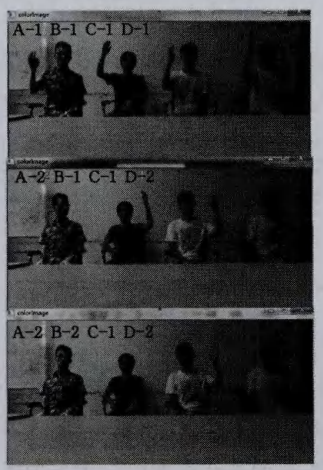


图 8 多学习者姿态识别的实验结果

从图 8 中可以看到，本文提出的方法能够正确识别出 4 位学习者的举手姿态和正坐姿态。

结束语 在数字化学习应用中，学习者的姿态信息对于学习者状态分析和行为记录具有重要的参考价值。目前利用教学场景中的学习者上半身的坐姿（如坐姿，举手等）来判断学习者状态的研究较少。本文以此作为研究内容，具有一定新颖性。

在室内人体姿态检测方面，目前研究比较多的是基于 Kinect 提供的人体骨骼信息的方法。该方法依赖于深度图像的二次处理，并且在识别人数上受到了 Kinect 可检测骨骼的人数限制。本文提出了一种根据深度图像对多学习者情况下的人体姿态进行估计的方法。该方法直接利用 Kinect 提供的深度图像，通过深度信息提取人像，克服了传统人像-背景方法鲁棒性不强的问题。在提取到完整的人体上半身轮廓之后，本文提出利用 Hu 不变矩和 SVM 分类器对学习者的坐姿进行估计。实验表明，这种方法从采集到的多学习者视频中可以获得对正坐、举手和低头等 3 种坐姿的较为准确的识别率。

本文对多学习者环境下的坐姿估计开展了有意义的尝试，在后续的工作中将对姿态库进行扩展和训练以适应更加广泛的应用场景。

参 考 文 献

[1] Ekman P, Bartlett M S, Hagger J C, et al. Measuring facial expressions by computer image Analysis[J]. Psychophysiology, 1999, 36(2): 253-263

[2] Wang H, Zhou L B, Ying Y. A Novel Approach for Real Time Eye State Detection in Fatigue Awareness System[C]// IEEE Conference on Robotics Automation and Mechatronics(RAM). 2010: 528-532

[3] Moeslund T B, Hilton A, Kruger V. A survey of advances in vision-based human motion capture and analysis[J]. Computer vision and image understanding, 2006, 104(2/3): 90-126

[4] 任海兵，徐光祐. 人体上肢姿态的估计及多解分析[J]. 软件学报，2002，13(11)：2127-2133
Ren Hai-bing, Xu Guang-you. Human Upper-Limb Pose Estimation and Multi-Resolution Analysis[J]. Journal of Software, 2002, 13(11): 2127-2133

[5] Agrarwal A, Triggs B. Recovering 3D human pose from monocular images[J]. IEEE Transactions on PAMI, 2006, 28(1): 44-58

[6] Hariharan B, Gopalakrishnan U. Gesture recognition using Kinect in a virtual classroom environment[C]// 2014 4th International Conference on Digital Information and Communication Technology and it's Applications. 2014: 118-124

[7] Hsu Hui-huang, Yi Chi-ou. Using Kinect to Develop a Smart Meeting Room[C]// 2013 16th International Conference on Network-Based Information Systems. 2013

[8] Vermun K, Senapaty M. Gesture-based Affective and Cognitive States Recognition using Kinect for Effective Feedback during E-learning[C]// 2013 IEEE 5th International Conference on Technology for Education. 2013: 107-110

[9] Hu M K. Visual Pattern Recognition by Moment Invariants[J]. IEEE Transactions on Information Theory, 1962(8): 179-187

[10] Huang Z H, Leng J S. Analysis of Hu's moment invariants on image scaling and rotation[C]// Proc. of the 2nd International Conference on Computer Engineering and Technology. 2010: 476-480

[11] Knerr S, Personnaz L, Dreyfus G. Single-layer learning revisited: A stepwise procedure for building and training a neural network[M]// Neurocomputing. Springer, 1990: 41-50