基于视角无关转换的深度摄像机定位技术*

韩云[1)] 钟圣伦[2)] 叶正圣[3)] 陈启军[1)†]

1) (同济大学电子与信息工程学院, 上海 201804)

2) (台湾科技大学电机工程系, 台北 10607)

3) (铭传大学计算机与通信工程学院, 台北 150001)

(2013 年 10 月 23 日收到; 2013 年 11 月 27 日收到修改稿)

通过整合深度和颜色信息, 深度摄像机 Kinect 能够稳健的侦测出人体及人体骨架关节点, 为计算机视觉、人体行为识别、机器人学的发展带来了革命性的进步. 然而单台深度摄像机的侦测范围有限. 虽然采用多台深度摄像机所构建的摄像机网可有效的扩大侦测范围, 但是必须依赖深度摄影机之间的相对位置与朝向的精确标定. 论文采用作者之前提出的以人体骨架为基础的视角无关转换技术, 能快速稳健的标定出深度摄像机之间的位置关系. 通过利用相邻两台深度摄影机同时侦测到的人体骨架, 论文能直接利用深度摄影机所量测的人体上半身中稳定的关节点为新坐标系的参考点, 实时的计算出两摄影机之间的平移向量和旋转矩阵, 而不依赖其他额外的校正设备或人为介入处理. 通过在室内环境中安装两台摆放于不同位置与朝向的深度摄影机, 从而, 验证了该方法的实时性与易用性. 该实时标定方法解决了深度摄影机侦测范围有限的限制, 同时, 可由两两相邻的标定扩展到多台深度相机的全局标定, 从而, 可以被广泛的应用于人体行为识别、情境感知服务等领域.

关键词: 深度摄像机网, 全局标定, 视角无关, Kinect

PACS: 42.30.Tz **DOI:** 10.7498/aps.63.074211

1 引 言

随着经济和社会的发展, 人体动作识别具有越来越广泛的应用前景, 如智能视频监控、人机交互、视频检索、体育训练、娱乐休闲等. 传统基于 2D 平面摄像机 [1,2] 的方法由于缺乏深度信息, 注定在前景识别中存在若干不可克服的困难. 随着技术的进步, 广大研究人员越来越意识到 3D 信息的重要性. 特别是微软推出性价比高的深度摄像机 Kinect 后, 采用深度影像技术极大的推动了动作识别的发展 [3−5]. 特别是与 Kinect 配套的中间件能够提供人体及骨架关节点的实时侦测, 为实时人体动作识别的发展带来了新的机会. 虽然 Kinect 在动作识别领域具有诸多优势, 但它存在一个较大的缺陷, 即 Kinect 有效侦测深度的范围大约是 0.5 m 到 4.5 m, 一旦超过该范围, 测量的误差将急剧增大. 故单个 Kinect 并不适用于侦测较大范围内的人体动作. 为了弥补 Kinect 的缺陷、扩大其侦测范围, 构建由多台 Kinect 组成的深度摄像机网就十分必要了. 深度摄像机网不仅可以扩大摄像机的侦测范围, 还能侦测到更加精细的动作, 为有效的识别人体动作打下坚实的基础. 图 1 是单台 Kinect 侦测的范围, 图 2 是两台 Kinect 侦测的范围. 从两幅图中, 可以很直观的看出摄像机网在扩大侦测范围上的优势.

整合两台及以上的摄像机来构建摄像机网, 关键的一步就是摄像机网的标定. 摄像机的标定可分为两类: 一类是单摄像机的标定; 另一类是多摄像机的全局标定 (global calibration). 单摄像机的

* 科技部国际合作项目 (批准号: 2010DFA12210)、上海科技人才项目 (批准号: 11XD1404800) 和上海科委科学基础研究重点项目 (批准号: 12JC1408800) 资助的课题.

† 通讯作者. E-mail: qjchenrobotics@gmail.com

标定是指求解出每台摄像机内外参数以及摄像机之间的关系. 常用的方法是Zhang提出的棋盘标定法[6]和Tsai提出的RAC 两部标定法[7]. 其中, Zhang提出的棋盘标定法, 借助由人为标记棋盘黑白棋格的顶点作为标记点, 标定精度高, 已成为一种主流的方法. 多摄像机的全局标定是指将所有摄像机的数据统一到总体的基准坐标系中, 即求出摄像机间的相对坐标, 也就是摄像机坐标系之间的平移向量T和旋转矩阵R. 核心思想是找出摄像机间共同侦测到的3D标记点来计算出摄像机之间的位置关系. 如贾倩倩采用液晶显示器上显示的图像序列为靶标来计算[8], Zhang采用的是一根杆上3个以上的标志点来计算的[9]. Thomas采用Kinect侦测到的点云为标记点, 借助ICP方法来计算的[10]. 贾倩倩和Zhang都采用的是2D平面摄像机, 而Thomas采用的是3D深度摄像机, 所以方法略有不同. 但是上述方法的共同缺点是需要额外准备特定的标记物, 并且计算量较大.

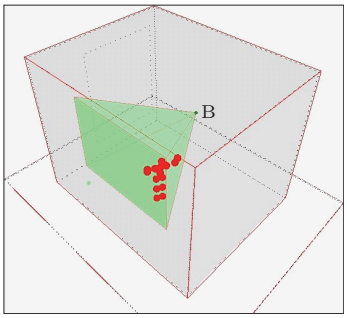

图1 一台Kinect侦测范围

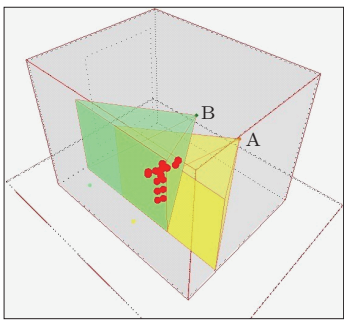

图2 两台Kinect侦测范围

为了构建深度摄像机网, 扩大深度摄像机的侦测范围, 以便整合多台深度摄像机来获取信息, 论文提出快速构建摄像网的标定方法, 即快速求解两台Kinect之间的平移向量T和旋转矩阵R的方法. 通过Kinect 侦测出的人体骨架中最稳定的由左肩与右肩所构成的肩轴与躯心 (torso center) 来构建人体骨架坐标系, 同时, 分别计算出两台摄像机相对人体骨架坐标系的平移向量和旋转矩阵, 然后以人体骨架坐标系为桥梁计算出两台摄像机间相对的平移向量和旋转矩阵, 从而, 实现将两台摄像机获取的信息统一到同一个坐标系统下的目的①. 该算法的早期研究结果见文献[11].

2 多台摄像机的标定技术与深度摄像机

2.1 多摄像机的信息融合基础

多台摄像机数据融合的基本思路是将所有摄像机侦测到的数据转换到某个基准坐标系下, 并经过一系列的整合形成一个整体. 其中的一个关键就是坐标系的转换. 以两台Kinect摄像机为例, 如图3所示. 每台Kinect都有自身的坐标系, 要实现两台摄像机侦测数据的融合, 第一步就是要选定其中一台摄像机的坐标系为基准坐标系, 将另一台摄像机侦测到的数据投影到基准坐标系中. 坐标系间的转换公式是

$$[P]_{\mathrm{A}} = R_{\mathrm{A}\leftarrow\mathrm{B}}\,[P]_{\mathrm{B}} + t_{\mathrm{A}\leftarrow\mathrm{B}}, \tag{1}$$

其中, $[P]_{\mathrm{A}}$表示空间中的一点P在A坐标系中的向量, $[P]_{\mathrm{B}}$表示同一点P在B坐标系中的向量, $R_{\mathrm{A}\leftarrow\mathrm{B}}$表示B坐标系相对A坐标系的旋转矩阵, $t_{\mathrm{A}\leftarrow\mathrm{B}}$表示B坐标系相对于A坐标系的平移向量, 整个公式表达的含义是将将同一向量点P, 从B坐标系中的表示, 转换到A坐标系中来表示. 即将B坐标系中的数据投影到A 坐标系中.

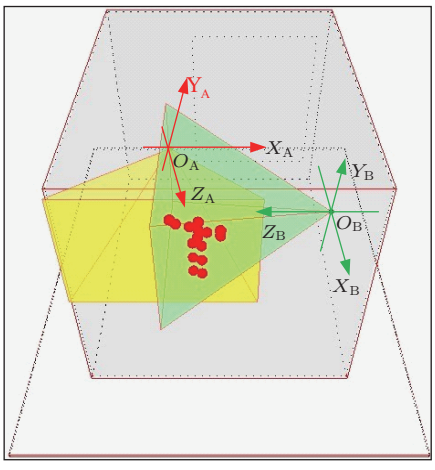

图3 两台摄像机间的坐标系转换

① 该研究的部分内容曾发表在 *IEEE International Conference on Systems, Man, and Cybernetics 2013* 上.

2.2 传统的求解旋转矩阵 R 与平移向量 T 的方法

由于 R 和 T 中一共有 4 组变量, 对应于每个方程是 4 个变量, 所以求解旋转矩阵 R 和平移向量 T 的思路是很清楚的, 即在空间中找到 4 个不同位置的 3D 点, 将其坐标值代入方程 (1), 就可求解出 R 和 T. 所以, 传统方法的思路是: 首先选定空间中的标记点, 然后找出 4 个以上标记点的 3D 坐标, 其次是代入 (1) 式求解出初步的 R 和 T, 最后进行优化, 求解出精确的 R 和 T. 不同方法的主要区别是标记点的选择不同, 如贾倩倩在 2D 摄像机下采用的是利用人工构建的视频序列, Mi 采用的是自由移动的平面模板 [12]. Zhang 采用的是带标号的棋盘格 [13]. 该类方法或者需要特定的辅助工具或者计算量较大, 都有待进一步的改进.

2.3 深度摄像机的视角规格与截头图

获取 3D 空间信息的想法由来已久, 最早是由 Marr 等提出的基于人类视觉的计算机视觉模型 [14] 来获取空间的三维信息 [15], 产生了双目立体摄像机系统. 后来陆续采用二维激光测距仪 [16]、三维激光测距仪 [17] 来获取 3D 空间信息. 但经济实用的深度摄像机却是微软公司于 2010 年发布的 Kinect 深度摄像机 [18]. 它具有性价比高的优势, 可以直接获得空间物体的 3D 信息, 并配置了强大的开发工具包. Kinect 有效侦测深度的范围是 0.5 m 到 4.5 m, 最佳侦测范围是 1.2 m 到 3.5 m, 水平视野是 57°, 垂直视野是 43°, 实体倾斜范围是 ±27°.

2.4 人体骨架侦测技术

深度摄像机在场景侦测上最大的优势是能够实时的提供人体骨架关节点的 3D 空间信息. 其实, 人体骨架提取的研究由来已久, 最初的研究主要是来自于人体姿态估计的需要. 如 Moeslund 和 Poppe 描述了姿态估计的发展 [19,20]. 后来, 人体关节点成为研究行为识别的一个重要特征. 但是由于多年来并不能实时、经济、精确的提取骨架信息, 而之前的许多研究都是假定能获取精确骨架这一前提下开展的, 所以许多基于骨架的动作识别算法并不具有真正的实用价值. 目前基于 Kinect 提出的实用人体骨架提取方法主要有两种: 一种是微软公司提出的方法 [21]; 另一种是 PrimeSense 公司在 OpenNI 库中提出的方法 [22]. Shotton 详细描述了微软公司的方法, 该方法的核心是结合机器学习中的随机决策森林和中介身体局部表示 (intermediate body parts representation) 来分类出每个身体部分, 然后采用基于分数可信度的方法, 确定出每个关节点的位置及姿态 [21]. OpenNI 提出的方法与微软的方法有许多相似的地方, 即都是在单幅深度图像的基础上提取人体骨架. 该方法理论上能获取 24 个关节点, 但目前真正有意义的只有 15 个, 详细见图 4. 具体关节名称见文献 [22]. OpenNI 骨架提取的过程可简单概括为: 深度产生器获取人体深度信息并传递给用户生成器, 用户生成器发现人体并进行标准姿势检测, 当检测到标准姿势时就进行骨架标定, 当标定完成后就进行骨架跟踪. 论文采用的实时快速标定方法是利用 Kinect 所提取人体骨架中最稳定的左肩、右肩和躯心 (torso center) 的 3D 空间位置信息来实现的.

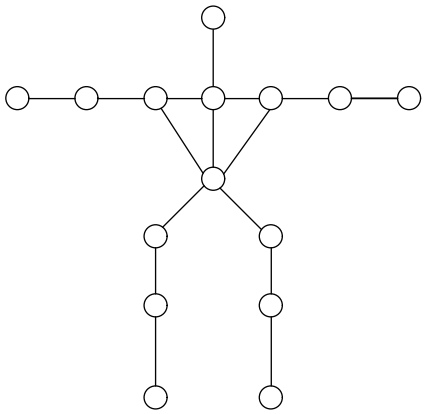

图 4 OpenNI 获取的人体骨架示意图

3 视角无关的摄像机网标定方法

在以比对为基础的动作识别中, 一个主要的障碍是: 作为分类依据的模板数据库获取图像时的摄像机位置与实际测试时的摄像机位置不一定在相同的相对位置上, 也就是说比对模板的视角和测试样本的视角可能不一致. 视角不一致给以比对为基础的动作识别带来了巨大的挑战. 论文定义“视角无关”为: 无论摄像机摆放的位置与视角如何, 都能将人体的骨架肢体与视角无关的最重要信息提取出来. 简言之, 就是将所有关节点以人体相对稳定的肩轴和躯心所构建的坐标系来表达. 同时, 结合坐标系转换来实现快速稳定的摄像机网标定.

3.1 以肩轴和躯心为基础的视角无关转换技术

深度摄像机Kinect在中间件OpenNI的协助下, 可有效获取人体的15个关节点的3D坐标及方向, 见图4; 使用Microsoft SDK则可获取20个关节点, 参见图5. 通过长时间的观察和实验发现, 在图4或图5的骨架关节点图中, 最稳定的关节点主要是上半身躯干的左肩、右肩、头和躯心这4个关节点. 同时, 还发现即使人在运动时, 上半身的左肩、右肩和躯心的相对位置也是很稳定的, 几乎保持不变, Kinect侦测出的结果也验证了这一发现, 而且以左肩与右肩所构成的肩轴, 取其中点和躯心的连线绝大部份时间都与肩轴垂直. 所以, 论文提出以左肩、右肩和躯心为基础构建描述身体自身的坐标系. 以左肩到右肩的方向为X轴方向, 以躯心到左右肩的中心点为Y轴方向, 通过右手定则来确定Z轴方向, 具体见图6. 论文将以人体上半身的这三点来建构新的坐标系, 并将原先由Kinect所量测的关节点转换到新的坐标系上的过程称为基于人体骨架的视角无关转换.

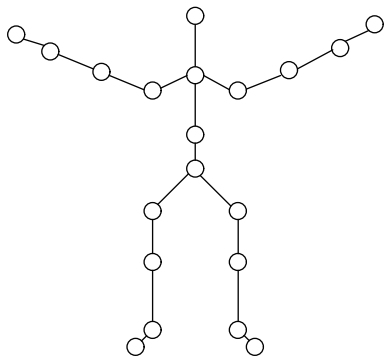

图5 Microsoft获取的人体骨架示意图

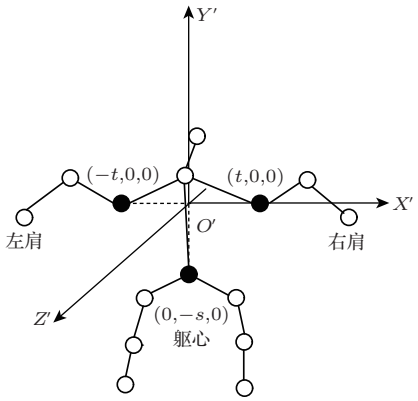

图6 以稳健关节点为基础构建的坐标系

将(1)式稍微修改一下, 得到描述摄像机坐标系与人体骨架坐标系关系的公式如下:

$$[P]_{\mathrm{ca}} = R_{\mathrm{ca}\leftarrow\mathrm{s}}\,[P]_{\mathrm{s}} + t_{\mathrm{ca}\leftarrow\mathrm{s}}, \tag{2}$$

与(1)式类似, $[P]_{\mathrm{ca}}$表示点P在摄像机坐标系下的表达, $[P]_{\mathrm{s}}$表示点P在人体骨架坐标系下的表达, $R_{\mathrm{ca}\leftarrow\mathrm{s}}$表示从人体骨架坐标系到摄像机间的旋转矩阵, $t_{\mathrm{ca}\leftarrow\mathrm{s}}$表示从人体骨架坐标系到摄像机间的平移向量.

参照(2)式, 考虑一台摄像机的坐标系以及新建的骨架坐标系, 如果减掉平移向量$t_{\mathrm{ca}\leftarrow\mathrm{s}}$, 剩余的即为旋转矩阵$R_{\mathrm{ca}\leftarrow\mathrm{s}}$, 其为线性运算. 按线性代数中所定义的线性转换的矩阵表示法(standard matrix representation of linear transformation), 只要找得到人体坐标系统中X, Y, Z轴的单位向量在Kinect摄像机坐标系中的表达向量, 即可依序求得旋转矩阵$R_{\mathrm{ca}\leftarrow\mathrm{s}}$中的三个竖直向量. 下面将介绍旋转矩阵的求法, 简言之, X, Y轴的表达式可以直接由Kinect的量测值取得; 相对应Z轴的量测值虽然不存在, 但可由所取的坐标系是正交的, 其相对应旋转矩阵也为正交, 所以, 可由前两个竖列向量通过外积法求得.

同时, 将(2)式右边的平移向量左移, 可得

$$[P]_{\mathrm{ca}} - t_{\mathrm{ca}\leftarrow\mathrm{s}} = R_{\mathrm{ca}\leftarrow\mathrm{s}}\,[P]_{\mathrm{s}}\,. \tag{3}$$

假定: $P = [P]_{\mathrm{ca}} - t_{\mathrm{ca}\leftarrow\mathrm{s}}$, 则(3)式表示为

$$P = R_{\mathrm{ca}\leftarrow\mathrm{s}}\,[P]_{\mathrm{s}}\,, \tag{4}$$

其中$R_{\mathrm{ca}\leftarrow\mathrm{s}}$可表示为

$$R_{\mathrm{ca}\leftarrow\mathrm{s}} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix},$$

如果

$$[P]_{\mathrm{s}} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^{\mathrm{T}},$$

则

$$R_{\mathrm{ca}\leftarrow\mathrm{s}}\,[P]_{\mathrm{s}} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

如果将右肩所在新坐标系的位置表示为$(t,0,0)$, 则可得

$$P = R_{ca\leftarrow s}[P]_{\mathrm{s}} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \begin{bmatrix} t \\ 0 \\ 0 \end{bmatrix} = t \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

即

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = P/t. \tag{5}$$

同理, 如果 $[P]_{\rm s} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^{\rm T}$, 则

$$R_{\rm ca\leftarrow s}[P]_{\rm s} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} d \\ e \\ f \end{bmatrix},$$

如果将躯心所在新坐标系的位置表示为 $(0, -s, 0)$, 则

$$P = R_{\rm ca\leftarrow s}\,[P]_{\rm s} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \begin{bmatrix} 0 \\ -s \\ 0 \end{bmatrix} = (-s) \begin{bmatrix} d \\ e \\ f \end{bmatrix},$$

可得

$$\begin{bmatrix} d \\ e \\ f \end{bmatrix} = P/(-s). \tag{6}$$

由右手定则, 可得 $R_{\rm ca\leftarrow s}$ 的第 3 列正好是第 1 列和第 2 列的外积, 则

$$\begin{bmatrix} g \\ h \\ i \end{bmatrix} = \begin{bmatrix} b \times f - e \times c \\ c \times d - f \times a \\ a \times e - d \times b \end{bmatrix}.$$

由于, 目前需要的是将摄像机坐标系映射到新建立的人体骨架坐标系中的旋转矩阵 $R_{\rm s\leftarrow ca}$, 由旋转矩阵的特点可得

$$R_{\rm s\leftarrow ca} = (R_{\rm ca\leftarrow s})^{\rm T}. \tag{7}$$

同理, 这里需要求解的是以新坐标系为基准的平移向量 $t_{\rm s\leftarrow ca}$. 由

$$[P]_{\rm s} = R_{\rm s\leftarrow ca}[P]_{\rm ca} + t_{\rm s\leftarrow ca}, \tag{8}$$

可得 $[P]_{\rm s} = R_{\rm s\leftarrow ca}\,[P]_{\rm ca} - R_{\rm s\leftarrow ca}t_{\rm ca\leftarrow s}$, 从而得到

$$t_{\rm s\leftarrow ca} = -R_{\rm s\leftarrow ca}t_{\rm ca\leftarrow s}. \tag{9}$$

3.2 坐标系间的转换

论文的研究目的是将 2 台摄像机的第 2 台 (非基准坐标) 所量测的点投影到第 1 台 (基准坐标) 中. 在上一步中, 求解了相对人体骨架坐标系的旋转矩阵和平移向量, 这一步的目的是求解出第 2 台摄像机相对第 1 台摄像机的旋转矩阵和平移向量. 思路是以人体骨架坐标系为桥梁来取得.

假定现在有三个坐标系: 摄像机 A 的坐标系、摄像机 B 的坐标系和人体骨架的坐标系. 由摄像机 A 到人的坐标系转换表示为

$$[P]_{\rm s} = R_{\rm s\leftarrow A}\,[P]_{\rm A} + t_{\rm s\leftarrow A}. \tag{10}$$

由摄像机 B 到人的坐标系转换表示为

$$[P]_{\rm s} = R_{\rm s\leftarrow B}\,[P]_{\rm B} + t_{\rm s\leftarrow B}. \tag{11}$$

因为

$$R_{\rm s\leftarrow A}\,[P]_{\rm A} + t_{\rm s\leftarrow A} = R_{\rm s\leftarrow B}P_{\rm B} + t_{\rm s\leftarrow B},$$

所以, 可得

$$R_{\rm s\leftarrow A}\,[P]_{\rm A} = R_{\rm s\leftarrow B}P_{\rm B} + t_{\rm s\leftarrow B} - t_{\rm s\leftarrow A}.$$

进一步, 可得

$$\begin{aligned} &(R_{\rm s\leftarrow A})^{-1}R_{\rm s\leftarrow A}\,[P]_{\rm A} \\ =&(R_{\rm s\leftarrow A})^{-1}(R_{\rm s\leftarrow B}\,[P]_{\rm B} + t_{\rm s\leftarrow B} - t_{\rm s\leftarrow A}), \\ &[P]_{\rm A} \\ =&(R_{\rm s\leftarrow A})^{-1}R_{\rm s\leftarrow B}[P]_{\rm B} \\ &+ (R_{\rm s\leftarrow A})^{-1}(t_{\rm s\leftarrow B} - t_{\rm s\leftarrow A}), \end{aligned}$$

所以

$$R_{\rm A\leftarrow B} = (R_{\rm s\leftarrow A})^{-1}R_{\rm s\leftarrow B}, \tag{12}$$

$$t_{\rm A\leftarrow B} = (R_{\rm s\leftarrow A})^{-1}(t_{\rm s\leftarrow B} - t_{\rm s\leftarrow A}). \tag{13}$$

最后, 将求解出的 $R_{\rm A\leftarrow B}$ 和 $t_{\rm A\leftarrow B}$, 以及非基准坐标系的中的数据代入 (1) 式, 即可将非基准坐标系中的数据映像到基准坐标系统中去.

4 多深度摄像机的标定实验

为了验证该算法在多摄像机标定上的有效性, 论文主要进行了两个方面的测试: 一是测试该标定算法的实时性与易用性; 二是测试该标定算法的精确度.

4.1 测试场景

为了测试算法的有效性, 选择了一间大约 20 $\rm m^2$ 的办公室, 安装了两台 Kinect 摄像机, 确保摄像机能侦测到整个空间. 整个环境及摄像机安放情况, 见图 7 和图 8.

硬件: 两台 Kinect for Windows、英特尔 Pentium G620@2.60 GHz 处理器、4 G 内存. 软件: 操作系统: windows7.0; 开发平台: vs2008; 附加库: OpenCV2.3.1, OpenNI2.2, Kinect for windows SDK v1.7.

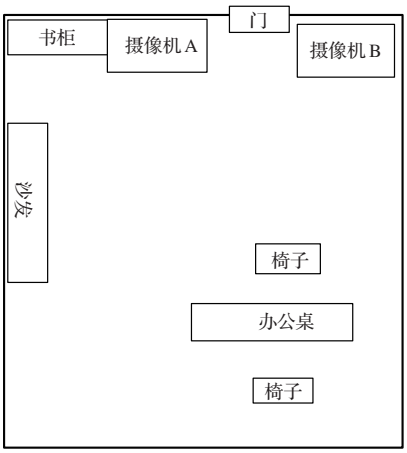

图 7 测试场景平面图 (摄像机并排)

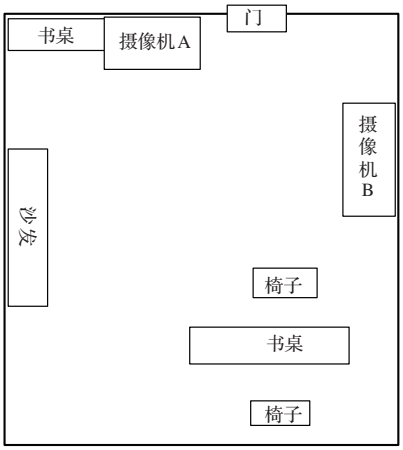

图 8 测试场景平面图 (摄像机夹角呈 90°)

4.2 标定的实时性与易用性测试及分析

为了验证算法的实时性与易用性, 论文设计了一个骨架配准实验. 实验的过程是: 假定目前有两台 Kinect 摄像机 A 和 B, 并选定 A 摄像机所在的坐标系为基准坐标系.

第一步: 两台 Kinect 同时获取同一个人体的骨架, 并显示出来.

第二步: 两台 Kinect 分别按上述提到的方法求解出摄像机坐标系相对人体骨架坐标系的旋转矩阵 $R_{\mathrm{s}\leftarrow\mathrm{A}}$, $R_{\mathrm{s}\leftarrow\mathrm{B}}$ 和平移向量 $t_{\mathrm{s}\leftarrow\mathrm{A}}$, $t_{\mathrm{s}\leftarrow\mathrm{B}}$.

第三步: 将旋转矩阵 $R_{\mathrm{s}\leftarrow\mathrm{A}}$, $R_{\mathrm{s}\leftarrow\mathrm{B}}$ 和平移向量 $t_{\mathrm{s}\leftarrow\mathrm{A}}$, $t_{\mathrm{s}\leftarrow\mathrm{B}}$ 代入 (11) 式和 (12) 式, 求出相对基准坐标系的旋转矩阵 R 和平移向量 T.

第四步: 将非基准坐标系所在摄像机 (这里指的是 B) 侦测到的骨架映射到基准坐标系 (这里指的是 A) 中, 并显示出来, 见图 9. 在该图中, (a), (b) 两图中的红点绿线骨架为两台 Kinect 独自获取的骨架图; 图 9 (a) 中的, 白点蓝线为将图 9 (b) 的骨架映射到图 9 (a) 中的效果. 通过观察基准坐标系中, 同一个人体由不同摄像机获取的骨架贴合度就能很直观的看出求解的 R 和 T 的效果.

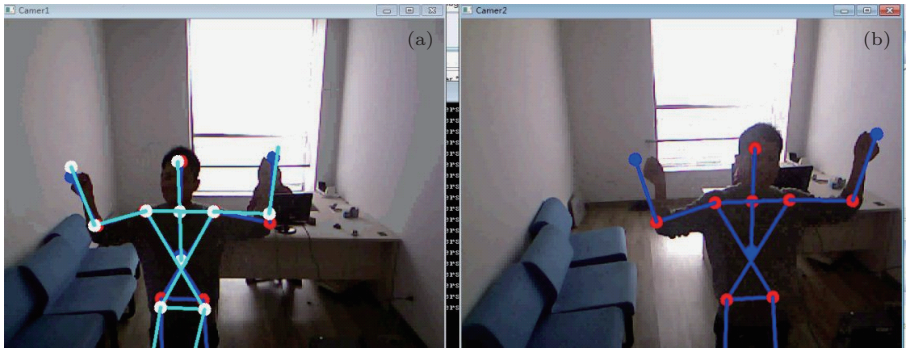

图 9 摄像机并排的骨架贴合图

为了验证算法的实时性, 在测试中, 每帧都计算一次 R 和 T, 整个过程是实时计算的. 目前, 求解 R 和 T 的时间包括两部分: 一是 Kinect 获取人体骨架的时间; 二是计算 R 和 T 的时间. 其中, Kinect 获取人体骨架的时间是程序初始化后, 侦测用户到姿势检测以及姿态标定等一系列过程的时间. 计算 R 和 T 的时间是获取骨架后, 计算整个 R 和 T 的时间. 通过实验得出的结论是: 获取骨架的平均时间是 3.5 s, 随着 Kinect 中间件技术的进步, 时间还会进一步的减少; 计算 R 和 T 的时间是 0.17 s, 所以第一次计算需要的时间大约是 3.67 s, 之后的计算只需要 0.17 s. 在论文中, 为了显示算法的稳健性才帧帧计算 R 和 T 的, 在实际全局标定摄像机网时, 只需取出代表性的几帧即可. 可见, 论文提出的算法是真正实时的.

为了进一步展示算法的易用性, 测试进行了两种摄像机的布局, 分别是: 两台摄像机并排、两台摄像机夹角呈 90°, 具体布局见图 7、图 8. 当摄像

机从图7的布局调整到图8的布局后, 只需要人在两个摄像机共同侦测到的区域站立一会即可完成标定. 而且, 从图9和图10可看出, 其骨架的贴合效果并没有因为摄像机的布局不同而发生变化. 所以, 从整个操作流程可看出: 该算法操作简单, 使用方便.

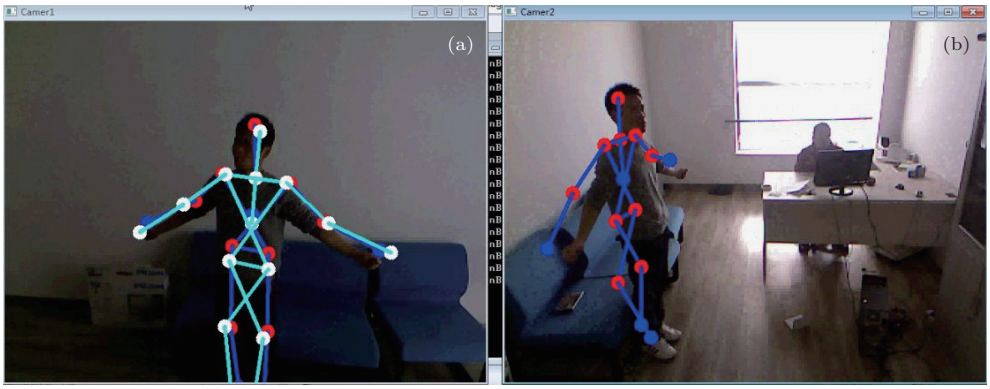

图 10 摄像机夹角呈 90° 的骨架贴合图

4.3 标定的精度评测及分析

为了测试算法的标定精度, 论文设计了在不同夹角下的比对实验, 即首先保证两个摄像机的朝向完全相同, 让其中一个摄像机固定不动, 另一个摄像机绕 Y 轴方向 (竖直方向) 旋转不同的角度, 然后通过人工测量两个摄像机间的夹角和距离 (即 $|t_{\text{真实}}|$), 同时, 记录下采用论文提出的算法求解出的两个摄像机间的夹角和距离 (即 $|t_{\text{量测}}|$), 最后, 比对人工测量的值与算法求解出的值, 从而得出算法的精确度. 其中, 平移向量的比较, 采用以下公式来表达, 即

$$\text{差值比例} = (|t_{\text{真实}}| - |t_{\text{量测}}|)/|t_{\text{真实}}|. \tag{14}$$

旋转矩阵的比较, 采用以下公式来表达, 即

$$\nabla R = \sqrt{([S]_{12})^2 + ([S]_{13})^2 + ([S]_{23})^2}. \tag{15}$$

其中, $S = R_{\text{真实}}(R_{\text{量测}})'$ 表示矩阵 $R_{\text{真实}}$ 和矩阵 $R_{\text{量测}}$ 的逆相乘. $[S]_{12}$ 表示 S 的第1行第2列的值.

实验一共测试了3种夹角下的情况, 分别是0°, 30°, 45°. 表1描述了算法量测的平移向量的精确度, 表2描述了算法量测的旋转矩阵的精确度.

表 1 摄像机不同夹角下 $|t|$ 的测试结果

测试角度 /(°)	0	30	45
$\lvert t_{\text{真实}}\rvert$/cm	110	153	153.5
$\lvert t_{\text{量测}}\rvert$/cm	111.12	149.02	149.25
相差/%	1.01	2.6	2.76

从表1可看出, 该算法在求解平移向量上的误差较小, 最大的只占到实际长度的2.76%, 仅仅相差几厘米. 从表2可看出, 该算法获取旋转矩阵的精确度较高, 最大误差仅0.1246, 大多数都小于0.1. 而在人体动作识别中, 人的肢体本身较大, 这些误差还远不足以对其造成影响. 同时, 还要看到该算法求解出的值与人工量测值之间还是存在一定误差的, 究其原因主要有两个方面: 一是由摄像机的

表 2 摄像机不同夹角下 R 的测试结果

测试角度 /(°)	$R_{\text{真实}}$	$R_{\text{量测}}$	∇R
0	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.9948 & -0.0996 & 0.0191 \\ 0.1007 & 0.9923 & 0.0708 \\ -0.0119 & -0.0724 & 0.9973 \end{bmatrix}$	0.1246
30	$\begin{bmatrix} 0.8660 & 0 & 0.5 \\ 0 & 1 & 0 \\ -0.5 & 0 & 0.8660 \end{bmatrix}$	$\begin{bmatrix} 0.8778 & 0.1732 & 0.4463 \\ -0.1384 & 0.9842 & 0.1098 \\ -0.4584 & -0.0346 & 0.8880 \end{bmatrix}$	0.0873
45	$\begin{bmatrix} 0.7071 & 0 & 0.7071 \\ 0 & 1 & 0 \\ -0.7071 & 0 & 0.7071 \end{bmatrix}$	$\begin{bmatrix} 0.7121 & 0.1016 & 0.6946 \\ -0.1071 & 0.9936 & 0.0355 \\ -0.6938 & 0.0491 & 0.7184 \end{bmatrix}$	0.0726

实际摆设以及人工量测误差引起的, 即在实际测试中, 由于人工摆设精度的原因导致摄像机不只是Y轴旋转, X轴和Z轴也存在略微的转动, 同时, 人工量测也有一定的误差. 二是由于Kinect获取人体骨架的精度引起的, 即目前Kinect获取人体骨架的精度还有待进一步的提高[23], 所以, 对该算法的精度也有一定的影响. 随着Kinect中间件技术的不断进步, 其侦测骨架的精度必将得到提升, 从而, 该算法的精度也将会得到相应的提高. 综上所述, 论文提出的算法完全能够满足动作识别的要求, 可广泛使用.

4.4 摄像机网带来的直接效果

为了更直观的展示由多台Kinect(这里以2台Kinect为例)构成的深度摄像机网的侦测效果, 论文通过比对一台摄像机与两台摄影机的有效监测范围来论述. 图1是一台Kinect侦测的有效范围大约是11.7 m^3. 图2是两台Kinect侦测的范围大约是21 m^3. 从2副图中可明显看出摄像机网侦测范围扩大了大约74%. 如果采用多台Kinect将能侦测更大的范围, 从而, 为室内动作的实时侦测打下坚实的基础.

5 结论与未来研究

在论文中, 提出了采用稳定的人体骨架关节点为参考点来构建视角无关的人体骨架坐标系, 并利用新提出的旋转矩阵和平移向量求解方法, 实现了基于多Kinect的深度摄像机网的实时标定. 论文的主要贡献是提出在空间上随意架设两台Kinect, 不需要额外的道具, 即可稳健实时的计算出其相对的旋转矩阵R与平移向量T, 从而, 实现深度摄像机网的构建.

在未来, 打算在两个方面开展进一步的研究: 一是深度摄像机网的构建与开发环境的数据融合; 二是室内较大范围内的人体动作识别. 通过采用多台深度摄像机实时构建体感摄像机网, 并快速实现数据的融合, 并以此为基础实现室内大场景的人体动作的实时监控与识别.

参考文献

[1] Xu Y N, Zhao Y, Liu L P, Zhang Y, Sun X D 2010 *Acta Phys. Sin.* **59** 980 (in Chinese) [许元男, 赵远, 刘丽萍, 张宇, 孙秀冬 2010 物理学报 **59** 980]

[2] Zhang W, Cheng B, Zhang B 2012 *Acta Phys. Sin.* **61** 060701 (in Chinese)[张伟, 成波, 张波 2012 物理学报 **61** 060701]

[3] Sung J, Ponce C, Selman B, Saxena A 2012 *IEEE International Conference on Robotics and Automation* Saint Paul, USA, May 14–18 2012 p842

[4] Xia L, Chen C C, Aggarwal J K 2012 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* Providence, United states, June 16–21, 2012 p8

[5] Yun K, Honorio J, Chattopadhyay D, Berg T L, Samaras D 2012 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* Providence, United states, June 16–21, 2012 p28

[6] Zhang Z Y 1999 *Proceedings of the IEEE International Conference on Computer Vision* Kerkyra, Greece, September 20–27, 1999 p666

[7] Tsai R Y 1986 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* Miami Beach, United states, June 22–26, 1986 p364

[8] Jia Q Q, Wang B X, Shi H, Luo X Z 2009 *Journal of Tsinghua University* **49** 676 (in Chinese) [贾倩倩, 王伯雄, 史辉, 罗秀芝 2009 清华大学学报 **49** 676]

[9] Zhang Z Y 2004 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26** 892

[10] Thomas E, Andreas S 2012 *7th German Conference on Robotics*Munich, Germany, May 21–22 2012 p1

[11] Han Y, Chung S L, Yeh J S, Chen Q J 2013 *IEEE International Conference on Systems, Man, and Cybernetics* Manchester, UK, October 13–16, 2013 p1525

[12] Mi T, An P, Liu S X, Zhang Z Y 2008 *Journal of Image and Graphics* **13** 1921

[13] Zhang T N, Meng C N, Liu R B, Chang S J 2013 *Acta Phys. Sin.* **62** 134204 (in Chinese)[张太宁, 孟春宁, 刘润蓓, 常胜 2013 物理学报 **62** 134204]

[14] Grimson W E L 1985 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-7** 17

[15] Wang F, Zhao X, Yang Y, Fang Z L, Yuan X C 2012 *Acta Phys. Sin.* **61** 084212 (in Chinese)[王芳, 赵星, 杨勇, 方志良, 袁小聪 2012 物理学报 **61** 084212]

[16] Atiya S, Hager G D 1993 *IEEE Transactions on Robotics and Automation* **9** 785

[17] Hoppen P, Knieriemen T, von Puttkamer E 1990 *IEEE International Conference on Robotics and Automation* Cincinnati, USA, May 13–18 1990 p948

[18] Weiss A, Hirshberg D, Black M J 2011 *IEEE International Conference on Computer Vision* Barcelona, Spain, November 6–13, 2011 p1951

[19] Moeslund T B, Hilton A, Krüger V 2006 *Computer Vision and Image Understanding* **104** 90

[20] Poppe R 2007 *Computer Vision and Image Understanding* **108** 4

[21] Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A 2011 *IEEE Conference on Computer Vision and Pattern Recognition* Colorado Springs, United states, June 20–25, 2011 p1297

[22] PrimeSense www. primesense. com [June 5, 2013]

[23] Han J G, Shao L, Xu D, Shotton J 2013 *IEEE Transactions on Cybernetics* **43** 1318

Calibration of D-RGB camera networks by skeleton-based viewpoint invariance transformation*

Han Yun[1)] Chung Sheng-Luen[2)] Yeh Jeng-Sheng[3)] Chen Qi-Jun[1)†]

1) (*College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*)

2) (*Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei 10607, China*)

3) (*Department of Computer and Communication Engineering, Ming Chuan University, Taipei 150001, China*)

(Received 23 October 2013; revised manuscript received 27 November 2013)

Abstract

Combining depth information and color image, D-RGB cameras provide a ready detection of human and associated 3D skeleton joints data, facilitating, if not revolutionizing, conventional image centric researches in, among others, computer vision, surveillance, and human activity analysis. Applicability of a D-RBG camera, however, is restricted by its limited range of frustum of depth in the range of 0.8 to 4 meters. Although a D-RGB camera network, constructed by deployment of several D-RGB cameras at various locations, could extend the range of coverage, it requires precise localization of the camera network: relative location and orientation of neighboring cameras. By introducing a skeleton-based viewpoint invariant transformation (SVIT), which derives the relative location and orientation of a detected human's upper torso to a D-RGB camera, this paper presents a reliable automatic localization technique without the need for additional instrument or human intervention. By respectively applying SVIT to two neighboring D-RGB cameras on a commonly observed skeleton, the respective relative position and orientation of the detected human's skeleton for these two cameras can be obtained before being combined to yield the relative position and orientation of these two cameras, thus solving the localization problem. Experiments have been conducted in which two Kinects are situated with bearing differences of about 45 degrees and 90 degrees; the coverage can be extended by up to 70% with the installment of an additional Kinect. The same localization technique can be applied repeatedly to a larger number of D-RGB cameras, thus extending the applicability of D-RGB cameras to camera networks in making human behavior analysis and context-aware service in a larger surveillance area.

Keywords: D-RGB camera network, network localization, viewpoint invariance, Kinect

PACS: 42.30.Tz **DOI:** 10.7498/aps.63.074211

* Project supported by the International S&T Cooperation Projects of China (Grant No. 2010DFA12210), the Shanghai Science and Technology Talent Project, China (Grant No. 11XD1404800), and the Key Program for Basic Research of Science and Technology Commission Foundation of Shanghai, China (Grant No. 12JC1408800).

† Corresponding author. E-mail: qjchenrobotics@gmail.com