文章编号: 1004-4353(2018)01-0069-05

基于 Kinect 信息融合的移动平台目标定位算法研究

李思含, 罗凯, 金小峰*
(延边大学 工学院, 吉林 延吉 133002)

摘要: 针对场景的光照变化和遮挡、混响等因素对目标定位准确性和鲁棒性的影响,提出了一种基于 Kinect 音视频融合的目标定位方法. 在获取场景的颜色、深度和声源定位信息后,首先利用获取的深度信息剔除背景信息,然后分别对颜色、深度和声源定位的模型计算似然函数,最后融合上述 3 种似然函数,并在粒子滤波框架下实现目标定位. 实验结果表明,音视频信息融合的目标定位平均准确率达到 90.7%,相比于同一场景下独立使用视频和音频定位的准确率分别提高了 9.1%和 16.9%.

关键词: Kinect; 信息融合; 深度信息; 目标定位; 声源定位; 粒子滤波
中图分类号: TP391.41 **文献标识码**: A

Research on mobile platform target localization algorithm based on Kinect information fusion

LI Sihan, LUO Kai, JIN Xiaofeng*
(*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*)

Abstract: Aiming at the influence of illumination variation, occlusion and reverberation on the accuracy and robustness of object location, a new method based on Kinect audio-video fusion is proposed. After obtain the color, depth and sound source location information of the scene, firstly, the background information is eliminated by depth information. Secondly, the likelihood function is computed for the model of color, depth and sound source location. Finally, fused three likelihood functions and implemented target location under framework of the particle filter. The experimental results show that the average accuracy of proposed method reaches to 90.7%, in contrast with singly using video and sound source location methods, the proposed method increased accuracy by 9.1% and 16.9% respectively.

Keywords: Kinect; information fusion; depth information; target localization; sound source localization; particle filter

0 引言

近年来,机器人技术得到迅猛发展,在各个行业得到广泛应用,如何让机器人能与人协同工作具有重要的研究意义. 机器人目标定位与跟踪方法主要是通过传感器获取目标的特征,并将其作为目标定位与跟踪的关键依据. 现有的目标定位方法可以分为基于视频和基于音频的定位方法. 基于视频的目标定位方法主要有基于区域、基于特征和基于模版的方法[1],其典型的算法有 Mean-shift[2-3]、卡尔曼滤波[4]和粒子滤波[5-6]等. 针对传统的 Mean-shift 算法对空中运动目标定位时存在较大形变和被遮挡的问题,文献[7]提出了基于 Mean-shift 算法和归一化转动惯量特征

收稿日期: 2017-04-21 **基金项目**: 吉林省科技厅自然科学基金资助项目(20140101225JC)
* **通信作者**: 金小峰(1970—),男,教授,研究方向为智能信息处理.

的目标定位算法.针对粒子滤波算法中存在的计算量大以及粒子退化等问题,文献[8]提出了融合Mean-shift 算法的粒子滤波定位方法,将粒子传递后的位置均值偏移到目标匹配的位置,使粒子快速收敛,以此减弱了粒子的退化问题.音频定位方法主要采用了文献[9]提出的广义互相关时延估计的音频定位算法,这种方法能够很好地计算出目标的方位,并且具有计算量小、精度高的优点.但是无论采用基于视频还是基于音频的目标定位方法,均存在单一传感器特征鲁棒性差的问题,因此近年来多传感器信息融合的方法成为目标定位与跟踪的主流方法.文献[10]通过对音频和视频分别采用到达时延的方法和内核背景差分法提取了跟踪目标的音视频特征,然后经序贯蒙特卡罗技术对上述两种特征进行了决策级融合.文献[11]采用带权重的粒子滤波框架对说话人的音视频信息进行了特征级融合,该方法能够有效地跟踪场景中出现的每一个说话人.通过对文献[10]和[11]的分析发现,文献中提出的音视频信息融合的目标跟踪或定位方法在黑暗或复杂场景中无法满足精确定位目标的要求;基于此,本文提出了一种将深度视频、彩色视频和音频在粒子滤波框架下进行特征级融合的目标定位方法,并通过实验验证了本文方法的有效性.

1 基于视频的目标特征提取

1.1 基于深度信息的目标检测

在视觉目标定位时,环境信息变化是影响定位准确率的主要因素,对此利用 Kinect 传感器所获取的深度信息对背景信息进行初步剔除.首先,将当前彩色图像所对应的深度图进行二值化,再将得到的二值图进行腐蚀、膨胀操作得到掩膜图,最后将掩膜图与原彩色图像进行掩模运算,得到如图 1 所示的背景剔除结果.

图 1 利用深度信息的目标提取结果

从图 1 可以看出,剔除的背景区域(黑色区域),相当于减小了目标定位方位,这为后期的基于颜色的粒子滤波算法的运用提供了有利的条件.

1.2 彩色图像的目标特征提取

传统的视觉特征主要有颜色、纹理和边缘等特征,其中颜色直方图应用最为广泛,但其对于复杂场景下目标图像的区分性并不是很理想[12];因此,本文在视觉特征提取中采用目标区域的颜色布局描述符.该描述符是国际标准 MPEG7 中建议的一种描述符,表达了图像的颜色空间信息.

首先将彩色图像从 RGB 空间映射到 YC_bC_r 空间,并将其中一个通道进行 8×8 大小的图像分块;然后分别计算每个子块的均值并构成均值矩阵,矩阵中每个值代表所对应子块的主信息;最后对均值矩阵进行离散余弦变换,变换结果通过之字形扫描和量化后得到一个通道的完整描述.

按上述方法对 Y、C_b、C_r 3 个通道分别提取特征,得到 3 组颜色分布特征向量,其中 Y 通道取前 6 个分量,C_b、C_r 通道取前 3 个分量.将取出的分量组合成一个共有 12 个分量的特征向量,此特征向量即为颜色布局描述符.

2 基于声源定位的目标音频特征提取

本文采用的目标音频特征是通过广义互相关到达时延的声源定位算法所确定的目标声源方位角,其主要思想是:根据广义互相关函数求得音频信号到达不同麦克风的时间延迟,进而通过麦克风阵列的几何结构计算声源的方向角[13].该算法计算量小且易于实现,定位精度能够满足一般系统的要求.根据移动平台目标定位的一般情况,本文采用远场模型[9]进行声源定位,该模型中假设麦克风接收到的声音信号都是相互平行的(如图 2 所示).图 2 中 L 为麦克风之间的距离.

根据麦克风接收到声音信号的广义互相关函数,求得函数峰值所对应的时间,该时间即为声音到达不同麦克风的时间延迟.根据图 2 所示远场定位模型,得到双麦克风定位方位角 θ.对于图 3 所示的 Kinect 传感器中麦克风的分布特征,本文采用分组的方式,将麦克风阵列分为 4 组,即 (M_1,M_2),(M_1,M_3),(M_1,M_4),(M_2,M_4).根据远场声源模型可知每组定位结果应为同位角,因

此 Kinect 声源定位的最终结果为

$$\theta=\frac{1}{N}\sum_{i=1}^{N}\theta_i, \tag{1}$$

其中 N 为麦克风个数.

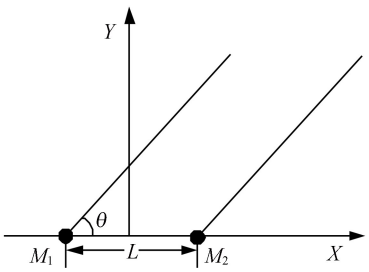


图 2　声源定位模型示意图

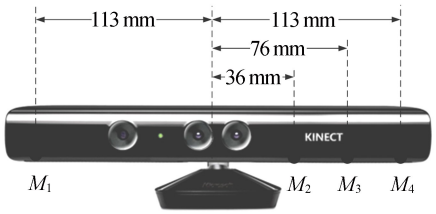


图 3　Kinect 麦克风阵列示意图

由于音频信号很容易受到环境噪声和混响的影响，需要对每一次定位的结果与上一次的定位结果做差分运算. 假设结果超出预设的阈值，则认为当前时刻的定位结果不可信，沿用上一时刻的定位结果.

3　基于粒子滤波的信息融合

3.1　构造似然函数

粒子与目标的相似度采用欧氏距离来度量. 此外，为了使粒子目标图像与模板图像的距离满足一定的分布，并且能呈现加权性质，即距离越近权值越大，反之越小，本文中似然函数采用高斯分布模型. 设彩色图像目标区域的测量方差为 σ_c^2，$d_i(x_i^k)$ 为第 k 时刻 i 粒子与目标的距离(相似度)，则颜色似然函数为

$$p(z_c^k \mid x_i^k)=\frac{1}{\sqrt{2\pi}\sigma_c^2}\exp\left(\frac{-d_i^2(x_i^k)}{2\sigma_c^2}\right). \tag{2}$$

假设深度视频的测量方差为 σ_d^2，k 时刻粒子集 $\{x_i^k\}$ 的深度似然函数为

$$p(z_d^k \mid x_i^k)=\frac{1}{\sqrt{2\pi}\sigma_d^2}\exp\left(\frac{-h_i^2(x_i^k)}{2\sigma_d^2}\right). \tag{3}$$

计算粒子和声源的方位角度差，需要确定粒子和目标的实际空间位置，即需要对 Kinect 进行标定[14]. 空间坐标系如图 4 所示，Kinect 放置于坐标原点，Z 轴正方向为声源定位方向.

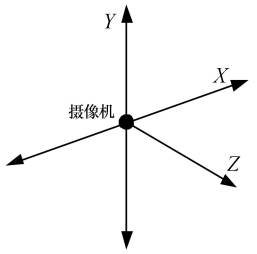


图 4　Kinect 空间坐标系

由于深度摄像头的成像模型与理想针孔模型十分相似，因此在精度允许的情况下，可以通过像素点偏离图像中心位置的坐标得到

$$\begin{cases} d_x=|x-x_0|\,\mathrm{depth}(x,y)/f, \\ d_y=|y-y_0|\,\mathrm{depth}(x,y)/f, \end{cases} \tag{4}$$

其中 d_x 和 d_y 表示像素点 (x,y) 偏离图像中心位置 (x_0,y_0) 在 X 和 Y 方向的偏移量，$\mathrm{depth}(x,y)$ 为该点对应的深度值，f 为摄像机的焦距. 假设粒子状态为 (u_i,v_i)，则粒子到麦克风阵列的角度为

$$\theta_i=\arctan(f/|u_i-x_0|), \tag{5}$$

假设 θ_t 为广义互相关时延估计的声源定位结果，由此得 k 时刻粒子集 $\{x_i^k\}$ 声源方位角似然函数为

$$p(z_a^k \mid x_i^k)=\frac{1}{\sqrt{2\pi}\sigma_a^2}\exp\left(\frac{-h_i^2(\phi_i^k)}{2\sigma_a^2}\right), \tag{6}$$

其中 $\phi_i^k=|\theta_i^k-\theta_t|$，声源定位结果方差为 σ_a^2.

综上，颜色、深度和声源方位角的似然函数都是基于高斯分布模型构造的，且三者的似然函数相互独立，因此对于独立同分布的联合似然函数可以表示为

$$\begin{aligned} p(z_k \mid x_i^k)&=C\cdot p(z_c^k \mid x_i^k)p(z_d^k \mid x_i^k)\cdot \\ p(z_a^k \mid x_i^k)&=\frac{C}{2\pi\sqrt{2\pi}\sigma_c^2\,\sigma_d^2\,\sigma_a^2}\cdot \\ &\exp\left\{\frac{-d_i^2(x_i^k)}{2\sigma_c^2}+\frac{-h_i^2(x_i^k)}{2\sigma_d^2}+\frac{-h_i^2(\phi_i^k)}{2\sigma_a^2}\right\}, \end{aligned} \tag{7}$$

其中 C 为调节常数，其作用是为了防止似然函数在更新过程中越来越小.

将式(7)求得的结果作为粒子滤波的传递权值，即可实现音视频信息的特征级融合. 假设颜色、深度和音频特征中有一个失效，则粒子的似然函数将分布在趋于零的位置，即权值极小但不为

零,这使得音视频融合的似然函数在某一传感器失效的情况下依然有效.

3.2 移动平台上的目标定位方法

本文提出的基于 Kinect 信息融合的移动平台目标定位算法的具体步骤如下:

Step 1 对检测到的目标的第一帧,选取目标特征区域,并计算该区域的颜色布局描述符 I_t.

Step 2 对深度图像进行归一化以及形态学操作,剔除大部分无关背景信息后进行下一步的精确定位.

Step 3 对粒子滤波参数进行初始化,粒子集为 $\{x_i^0\}_{i=1}^N$;然后通过状态转移模型进行粒子的传递,得到下一时刻的粒子集 $\{x_i^1, w_i^1\}_{i=1}^N$.

Step 4 根据式(2)、(3)和(6)计算粒子的彩色、深度和声源方位角的似然函数.

Step 5 根据式(7)计算音视频的融合似然函数,并计算粒子的最大后验概率所对应的粒子状态,该状态即为当前时刻的定位结果;同时对粒子分布进行评估,重采样后转 Step 2.

循环 Step 2—Step 5,即可实现目标的实时定位. 在定位过程中,将目标的定位结果反馈给移动平台,由此产生平台转向与移动的指令,实现平台持续对目标的定位与跟踪功能.

4 实验及分析

为了验证本文提出算法的有效性,进行似然函数的对比实验和音视频信息融合的目标定位实验. 传感器采用 Kinect(Windows 版),并将 Kinect 放置于移动平台上,实验过程中分别采集场景的深度、彩色和音频信息.

4.1 似然函数的对比实验

随机在目标区域中选取 50 个粒子,并计算粒子的视频颜色、音频(声源)方位以及音视频融合的似然函数. 似然函数值与粒子权值的对应关系如图 5 所示. 从图 5 可以看出,独立的视频颜色和音频似然函数计算的粒子权值呈现出众多虚假的目标定位结果,且两者定位不一致的情况,而音视频融合的方法可以保证高权值赋予那些视频颜色和音频似然函数值同时大的粒子,即视频中和音频中最接近目标的粒子被赋予了较大的权值,否则赋予较小的权值,这说明该方法提高了目标定位的准确性和鲁棒性. 但在实验过程中也发现,某些粒子由于受光照、遮挡、噪声和混响等因素的影响,使得粒子的权值过小,从而导致退化现象的发生.

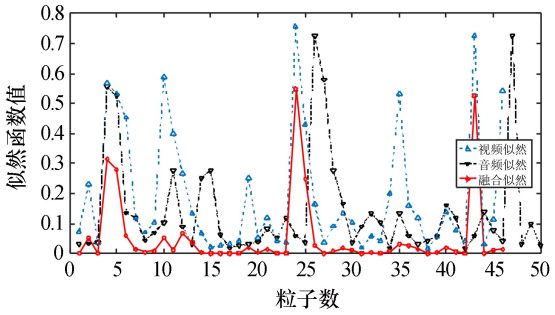
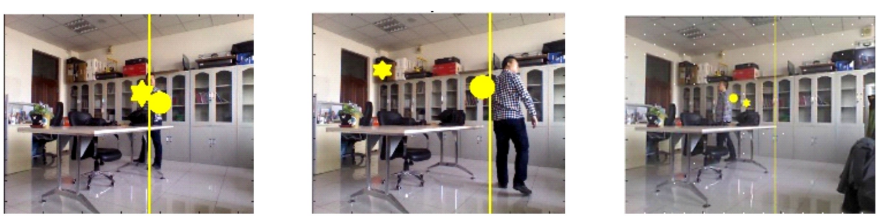


图 5 视频颜色、音频、音视频融合的似然函数对比图

4.2 音视频融合的目标定位实验

实验中随机选取了 4 组连续的 200～300 帧音视频采样数据,分别进行了单独的视频定位和音频定位以及音视频融合定位实验,定位结果部分截图如图 6 所示,实验数据见表 1.

在图 6 中,六角星表示单独视频的目标定位点,竖线表示单独音频的目标定位方位,实心圆表示音视频融合的目标定位点. 目标定位结果与目标矩形区域有交叠部分判定为定位准确,否则判定为定位错误. 图 6(a)是音视频融合定位准确的结果,即实心圆与定位目标吻合;图 6(b)是单独视频定位错误的结果,即六角形已严重偏离了目标;图 6(c)是单独音频定位错误的结果,黄色竖线偏离了目标. 但是,图 6 中表示音视频融合定位的结果均是准确的.



(a) 音视频准确定位 (b) 视频定位丢失 (c) 音频定位丢失

图 6 视频、音频、音视频融合定位的结果示意图

表 1　视频、音频、音视频融合的定位结果

实验	总帧数	视频单独定位的准确率/%	音频单独定位的准确率/%	音视频融合定位的准确率/%
1	212	84.9	72.2	92.0
2	265	74.0	69.8	89.4
3	249	77.1	80.7	88.4
4	286	90.6	72.4	93.0
平均准确率/%		81.7	73.8	90.7

从表 1 中可以看出，4 组实验音视频融合定位的平均准确率达到 90.7%，相比于同一场景下的单独视频定位和音频定位的准确率分别提高了 9.1%和 16.9%.

5　结论

本文提出的基于 Kinect 信息融合的移动平台目标定位算法，对光照变化、遮挡、噪声和混响等干扰因素具有较强的鲁棒性，目标未发出声音时可通过视频信息对其定位，而当目标突然离开视野时可通过音频信息确定目标的方位. 另外，当目标处于在黑暗且无声音的场景下，依然可以通过深度信息对运动目标定位. 本文方法在使用粒子滤波过程中仍然存在粒子退化和匮乏等问题，需要进一步深入研究以完善本文方法.

参考文献：

[1] Cheng C, Ansari R. Kernel particle for visual tracking[J]. IEEE Signal Processing Letters, 2005,12(3):242-245.

[2] 宋丹，赵保军，唐林波. 融合角点特征与颜色特征的 Mean-Shift 目标跟踪算法[J]. 系统工程与电子技术，2012,34(1):199-203.

[3] 张玲，蒋大永，何伟，等. 基于 Mean-shift 的改进目标跟踪算法[J]. 计算机应用，2008,28(12):3120-3122.

[4] 李剑汶，周慧，王小阳，等. 基于超声波时差定位和卡尔曼滤波的服务机器人导航方法[J]. 南京大学学报(自然科学版)，2015(S1):82-86.

[5] 闫河，刘婕，杨德红，等. 基于特征融合的粒子滤波目标跟踪新方法[J]. 光电子·激光，2014(10):1990-1999.

[6] 许婉君，侯志强，余旺盛，等. 一种改进的多特征融合目标跟踪算法[J]. 电光与控制，2015(12):34-39.

[7] 甘明刚，陈杰，王亚楠，等. 基于 Mean Shift 算法和 NMI 特征的目标跟踪算法研究[J]. 自动化学报，2010,36(9):1332-1336.

[8] 谢静. 基于音视频融合的目标跟踪算法[D]. 天津：天津大学，2009:47-55.

[9] 陶巍，刘建平，张一闻. 基于麦克风阵列的声源定位系统[J]. 计算机应用，2012,32(5):1457-1459.

[10] Hoseinnezhad R, Vo B N, Vo B T, et al. Bayesian integration of audio and visual information for multi-target tracking using a CB-member filter [C]//IEEE International Conference on Acoustics, Speech, & Signal Processing, 2011: 2300-2303.

[11] Steer M A, Al-Hamadi A, Michaelis B, et al. Audio-visual data fusion using a particle filter in the application of face recognition[C]//20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010:4392-4395.

[12] Kasutani E, Yamada A. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval [C]//International Conference on Image Processing, 2001:674-677.

[13] 景思源，冯西安，张亚辉. 广义互相关时延估计声定位算法研究[J]. 声学技术，2014(5):464-468.

[14] 郭连朋，陈向宁，刘彬. Kinect 传感器的彩色和深度相机标定[J]. 中国图像图形学报，2014,19(11):1584-1590.