doi: 10.12068/j.issn.1005-3026.2020.07.004

基于BA优化和KL散度的RGB-D SLAM系统

徐 岩, 安卫凤
(天津大学 电气自动化与信息工程学院, 天津 300072)

摘 要: 针对深度相机采集深度图像的噪声对位姿估计精度的影响,以及误差随时间累积的严重问题,设计了一种改进的基于RGB-D相机的视觉SLAM系统. 首先,建立重投影误差模型,通过最小化重投影误差,对位姿和特征点进行非线性优化. 此外,提出了一种闭环检测的算法,建立字典模型,用频率-逆文档频率计算权重,用Kullback-Leibler散度计算相似度,并使用相对相似度机制检测闭环,减少了累积误差. 使用15个公开的图像序列对算法进行评价,同3个流行的RGB-D SLAM系统对比,精度平均最高提高119.07%,最低提高4.24%. 实验结果证明,提出的方法比目前流行的RGB-D SLAM系统具有更好的精度.

关 键 词: 机器视觉;同时定位与建图;BA优化;KL散度;闭环检测
中图分类号: TP 242.6 **文献标志码**: A **文章编号**: 1005-3026(2020)07-0933-05

RGB-D SLAM System Based on BA Optimization and KL Divergence

XU Yan, *AN Wei-feng*
(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. Corresponding author: AN Wei-feng, E-mail: awf_joker@tju.edu.cn)

Abstract: To reduce the impact of noise in depth image acquisition on the accuracy of pose estimation and to solve the serious problem of cumulative error over time, an improved RGB-D SLAM system was designed. Firstly, the re-projection error model was established to nonlinearly optimize the poses and features by minimizing the re-projection error. In addition, a closed-loop detection algorithm was proposed. The dictionary model was established and the frequency-inverse document frequency(TF-IDF) was used to calculate the weight. Kullback-Leibler divergence was used to calculate the similarity, and a relative similarity mechanism was used for the closed-loop detection. The cumulative error was decreased. The algorithm was evaluated using 15 public image sequences. Compared with three popular RGB-D SLAM systems, the maximum increase of accuracy is averagely 119.07% and the minimum increase is 4.24%. Experimental results show that the proposed method has better accuracy than the current popular RGB-D SLAM systems.

Key words: machine vision; simultaneous localization and mapping; BA optimization; KL divergence; closed-loop detection

随着室内移动机器人的普及以及自动驾驶技术的发展,同时定位与建图(simultaneous localization and mapping,SLAM)受到广泛关注. SLAM通过前端估计相机的位姿和局部地图,通过后端对位姿和地图进行优化,建立全局一致的地图,通过闭环检测消除累积误差,提高定位与建图的精度.

SLAM主要通过滤波器法[1]或者图优化方法[2]建图. 滤波器方法主要包括信息滤波器[3]、迭代卡尔曼滤波器[4]、无迹卡尔曼滤波器[5]和粒子滤波器[6]等. 基于滤波器的方法假设了马尔可夫性,即认为当前时刻的状态只与前几个有限时刻的状态有关,而与很久之前时刻的状态无关,不利于进行闭环处理. 而且,滤波器方法存储的状态量呈平方增长,在大型场景中,定位精度低,不适合实时建图. 此外,滤波器方法用三维坐标点表示

收稿日期: 2019-05-24
基金项目: 国家自然科学基金资助项目(61632018); 青海省基础研究项目(2017-ZJ-753).
作者简介: 徐 岩(1977-),女,辽宁沈阳人,天津大学副教授.

路标,假设地标点和位姿满足某种概率分布,通过更新滤波器的方式使三维坐标点收敛,当路标点的位置移动时,估计的位姿具有较大的误差,不适用于动态场景. 图优化方法采用非线性优化模型,使用捆集调整(bundle adjustment,BA)[7]算法优化位姿和地图,考虑了更多的历史时刻信息,克服了滤波器方法只考虑有限时刻信息的缺点,可以进行闭环检测与闭环调整. 而且,BA 求解的过程中,方程组的系数矩阵会出现稀疏性结构,加速了求解过程,适用于大型场景,有利于实时实现 SLAM,成为 SLAM 的主流优化方法.

视觉 SLAM 中常用的传感器是单目[8]、双目[9-13]和深度相机[14]. 深度相机通过结构光或飞行时间的物理方法测量物体的深度,广泛应用于室内定位中,缺点是噪声大. RGBDTAM[15]通过多视觉约束和最小化半稠密光度误差来提高定位精度. 但是,通过三角化恢复的深度精度不高,而且需要满足光度不变性的假设,对于光照不均匀的场合,定位误差较大. ElasticFusion[16]利用迭代最近点(iterative closest point,ICP)[17]算法估计位姿,采用不断地优化和重建地图的方式,提高地图重建和位姿估计的精度. 但是,不适用于重建较大的场景. Endres 等[18]利用 ICP 估计相机位姿,建立 3D-3D 匹配误差模型,采用不断优化特征点的方式来提高位姿的精度. ICP 算法适用于深度已知的特征点,对深度误差大的特征点的位姿估计误差较大. Mur-Artal 等[19]提出的 ORB-SLAM2 在估计位姿时,建立 3D-2D 重投影误差模型,利用图优化的方式优化位姿. 该方法只进行了位姿优化,对深度信息没有进行优化.

闭环检测是视觉 SLAM 一个重要的环节,通过闭环调整可以消除累积误差. 词袋模型(bag of words,BoW)[20]把图像特征看作单词,利用字典查找对应的单词,广泛应用于闭环检测中. 词袋模型中所用的字典利用 K-means 算法离线训练得到,二进制字典的特征向量的权重用二进制表示,加载时间短,但是准确度不高. 频率-逆文档频率(term frequency-inverse document frequency,TF-IDF)[21]通常用在分类中计算特征点的权重,Kullback-Leibler(KL)散度[22]用来衡量两个概率分布的差距,计算文本内容的相似度,常应用于词汇语音识别[23]、面部表情识别[24]和手写字符识别[25].

针对上述 SLAM 系统存在的深度误差对定位精度造成影响的问题,本文对位姿和特征点同时进行非线性优化,建立 3D-2D 重投影误差模型,最小化重投影误差;在闭环检测中,针对二进制字典检测闭环精度低的问题,使用词袋模型,用频率-逆文档频率计算特征向量的权重,用 KL 散度进行相似性评分,并提出了一种计算相对相似度的机制,避免引入绝对得分与绝对阈值. 实验结果表明,提出的算法可以有效地减小累积误差,具有较高的定位精度.

1 系统设计

本文设计的系统分为4个模块,如图1所示,即视觉里程计、后端优化、闭环检测和建图. 视觉里程计负责相机位姿的粗略估计. 后端采用捆集调整优化算法优化局部地图和全局地图,建立重投影误差模型,优化相机的位姿和特征点,提高相机位姿估计的精度. 在闭环检测中,本系统使用 KL 散度计算关键帧的相似度,利用相对相似度机制提高闭环检测的精度,检测到闭环之后,进行闭环矫正,提高定位精度. 为保证 SLAM 的实时性,本文利用保存的稀疏点云建立稀疏地图.

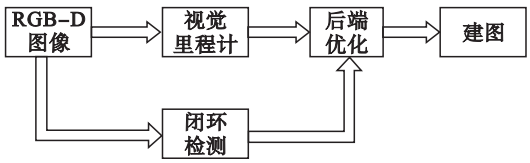

图1 系统框架

Fig. 1 System framework

1.1 优化模型的建立

为了减小深度误差对定位精度的影响,本文建立 3D-2D 重投影误差模型,使用少量的匹配点估计相机的位姿,通过最小化重投影误差,优化位姿和特征点. SLAM 的优化是求解非线性和非凸优化问题的过程.

空间点 $\boldsymbol{p}$,经内参矩阵为 $\boldsymbol{K}$、外参为($\boldsymbol{R}$,$\boldsymbol{t}$)的针孔相机投影,在像素平面上得到 $\boldsymbol{p}$ 的像素坐标为 z,根据针孔相机成像模型可知,z 与 $\boldsymbol{p}$ 满足:

$$sz=\boldsymbol{K}(\boldsymbol{Rp}+\boldsymbol{t})\ . \tag{1}$$

相应地,重投影误差为

$$\boldsymbol{e}=z-\frac{1}{s}\boldsymbol{K}\exp(\boldsymbol{\xi}^{\wedge})(\boldsymbol{p},1)^{\mathrm{T}}. \tag{2}$$

式中:s 为尺度因子;$\boldsymbol{\xi}^{\wedge}$ 为相机外参($\boldsymbol{R}$,$\boldsymbol{t}$)对应李代数的反对称矩阵;$(\boldsymbol{p},1)^{\mathrm{T}}$ 为空间点 $\boldsymbol{p}$ 的齐次坐标表示.

对重投影误差求和,构建最小二乘问题. 建立图优化模型对这个最小二乘问题求解,位姿和特征点待优化,实现位姿和特征点的非线性优化. 重

投影误差的目标函数为

$$\xi^* = \frac{1}{2}\sum_{i=1}^{n}\left\| z_i - \frac{1}{s_i}\boldsymbol{K}\exp(\boldsymbol{\xi}^\wedge)\boldsymbol{p}_i \right\|_2^2. \quad (3)$$

把特征点和位姿添加到图的顶点，重投影误差添加到边，通过 Levenberg－Marquardt 方法最小化重投影误差，从而优化位姿和特征点. 优化位姿和特征点增加了节点和约束边的数量，使得位姿更加精确.

1.2　基于 KL 散度相似性的闭环检测算法

二进制字典将特征权重设为 1 或者 0，表示单词在图像中出现与否，忽略了单词出现的数量，对于相似度高的环境，比如，办公室相似的桌椅，该方法不能准确地区分闭环. 提出的方法使用 TF－IDF 计算特征的权重，对重要度和区分度不同的特征给予不同的权重. KL 散度是一种信息熵，常用来衡量两个概率分布的差距，计算文本内容的相似度. 提出的方法利用 KL 散度计算图像的相似度，进行闭环检测.

设 η 是单词的权重，表示单词的重要性和区分性，使用 TF－IDF 计算，公式如下：

$$\eta = f_{\mathrm{TF}} \times f_{\mathrm{IDF}}. \quad (4)$$

其中：f_{TF} 表示特征在一幅图像中出现的频率；f_{IDF} 是叶子节点中特征数量占整个字典所有特征数量的比例，表示单词在字典中出现的频率. η 服从概率分布，用 KL 散度计算图像 A 对图像 B 的相似度：

$$s(A \parallel B) = \sum \eta_A \ln \frac{\eta_A}{\eta_B}. \quad (5)$$

相似度的分值是非负值，分值越低，表示两幅图像越相似，同一幅图像的分值为 0. 对于不同视角下场景变化较明显的环境，用 KL 散度计算出的相似度可以判断是否存在闭环. 但是对于不同视角下场景变化不明显的环境，分值的绝对大小并不具有绝对的代表性，设定固定的分数阈值并不能适应所有的环境.

针对这个问题，本文提出了一种计算相对相似度的机制，相对相似度定义为当前关键帧和以前某一时刻关键帧的相似度与当前关键帧和上一时刻关键帧的相似度的比值. 在闭环检测序列中，设 $t_1, t_2, \cdots, t_j$ 时刻对应的关键帧图像为 $I_{t_1}, I_{t_2}, \cdots, I_{t_j}$，当前时刻 $t_k\ (k>j)$ 的关键帧对 t_j 时刻的关键帧的相似度为 $s(I_{t_k} \parallel I_{t_j})$，$t_k$ 时刻的关键帧对 t_j 时刻的关键帧的相对相似度记为 $c(I_{t_k}, I_{t_j})$，按照式(6)计算.

$$c(I_{t_k}, I_{t_j}) = \frac{s(I_{t_k} \parallel I_{t_j})}{s(I_{t_k} \parallel I_{t_{k-1}})}. \quad (6)$$

计算新增关键帧与地图中的关键帧的相对相似度，选取最小的相对相似度作为动态的阈值，采取闭环帧缓存机制，若在一段时间内关键帧的相对相似度连续多次低于动态阈值，将其设为闭环候选帧，并更新动态阈值. 相对相似度机制利用相对相似度取代了绝对相似度，解决了相机在不同视角下场景变化不明显的环境中闭环检测不准确的问题，适用于多种环境.

检测到闭环帧后，利用空间一致性约束进行闭环验证. 闭环是相机出现在之前出现过的同一位置不同视角，满足对极几何约束，通过对相似帧进行特征匹配，利用随机抽样一致算法计算出两帧的基础矩阵，用这个基础矩阵计算内点，如果内点数多于先前设定的阈值，则认为是闭环. 计算出当前帧与闭环帧的相似变换[26]，进行闭环矫正. 闭环检测的流程如图 2 所示.

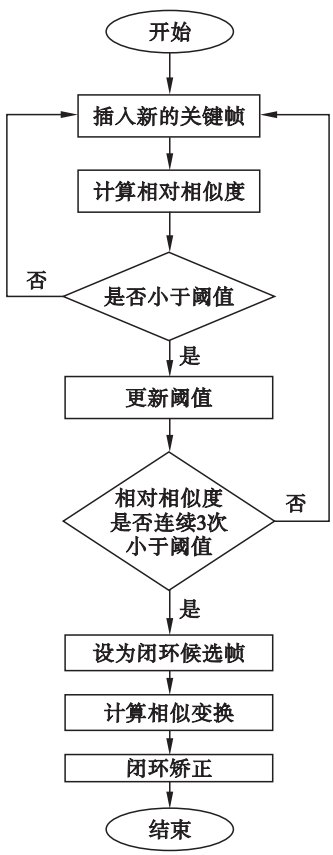

图 2　闭环检测流程

Fig. 2　Flow chart of closed-loop detection

2　实验结果与分析

为评价提出算法的性能，本文选用公开的 TUM RGB－D 数据集[27]进行实验，从定量角度比较本文提出的闭环检测算法与基于二进制字典的闭环检测算法的性能. 此外，与 3 个流行的

RGB－D SLAM 系统进行对比,定量地评价本文提出算法的总体效果.

为了比较提出的闭环检测算法与基于二进制字典的闭环检测算法的性能,将使用相对相似度机制的闭环检测和基于二进制字典的闭环检测作对比,使用序列 fr3_long_office_household 进行验证. 该序列由深度相机在办公室场景中沿着一个大的圆形移动拍摄得到,具有很多纹理和结构. 序列共 2 585 帧,轨迹长度为 21.455 m,时长 87.09 s. 轨迹的末端与开头重叠,存在一个大的、不规则的闭环. 实验结果用估计位姿的绝对误差表示,图 3 显示了用本文算法和用二进制字典估计位姿的绝对误差的直方图分布. 由图 3 可知,提出的闭环检测算法对相机位姿估计的绝对误差比使用二进制字典估计的绝对误差分布更集中,并且集中在误差较小的范围内.

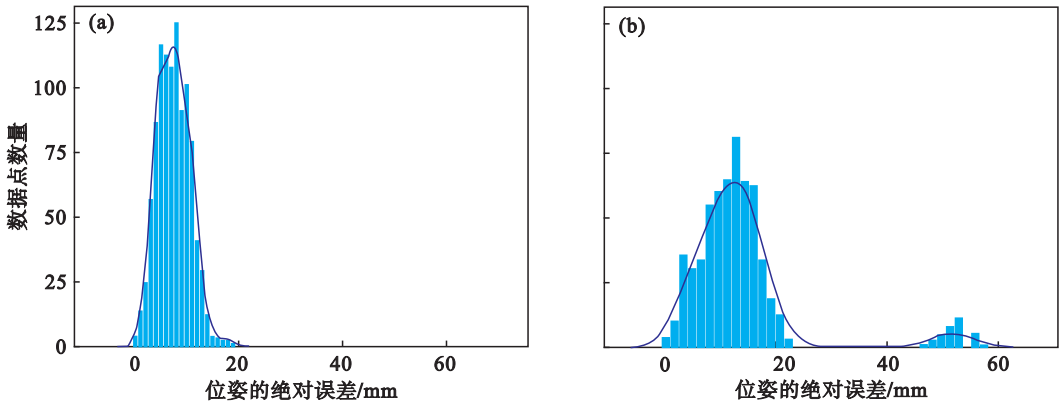

图 3 不同方法估计位姿的绝对误差直方图

Fig. 3 Histograms of absolute error in pose estimation by different methods

(a)—本文算法;(b)—二进制字典.

为了评价提出方法的总体效果,本文对比了 3 个流行的 RGB－D SLAM 系统,分别是 RGBDTAM[15]、ElasticFusion[16] 和 ORB－SLAM2[19]. 图像序列信息以及均方根误差如表 1 所示,“—”代表帧丢失,位姿无法恢复. 从表 1 可以看出,在 15 组实验数据中,有 8 组数据比 3 个系统均有不同程度的提高,尤其是第 2 组和第 7 组数据;由第 10 组和第 11 组数据可以看出,本文算法估计位姿的精度虽然比 RGBDTAM 估计的精度低,但是较 ORB－SLAM2 有明显提高;第 5,6 和 12 组数据显示,提出的算法位姿估计的精度虽然比 ORB－SLAM2 低,但是比 RGBDTAM 和 ElasticFusion 均有大幅度提高. 通过表 1 可以计算出本文方法对位姿估计的精度平均比 RGBDTAM,ElasticFusion 和 ORB－SLAM2 分别提高了 75.07%,119.7% 和 4.24%.

表 1 不同 SLAM 系统位姿估计均方根误差的比较

Table 1 Comparison of root mean square error of pose estimation for different SLAM systems m

序号	本文算法	RGBDTAM[15]	ElasticFusion[16]	ORB－SLAM2[19]
1	**0.015**	0.027	0.020	0.016
2	**0.018**	0.042	0.048	0.022
3	**0.010**	0.010	0.011	0.010
4	**0.021**	0.021	0.025	0.022
5	0.043	0.081	0.083	**0.041**
6	0.062	0.155	0.068	**0.047**
7	**0.008**	0.027	0.071	0.009
8	**0.004**	0.007	0.011	0.004
9	0.017	—	0.049	**0.013**
10	0.038	**0.026**	0.074	0.043
11	0.018	**0.010**	0.016	0.025
12	0.010	0.013	0.030	**0.009**
13	**0.020**	0.044	0.021	0.021
14	**0.010**	0.010	0.013	0.011
15	0.011	**0.010**	0.015	0.011

注:表中加黑数字代表本组实验的最好结果.

3　结　　语

本文提出了一种减小深度信息误差的方法，通过建立 3D – 2D 重投影误差优化模型，用 Levenberg – Marquardt 方法最小化重投影误差，优化特征点和位姿. 此外，提出了一种闭环检测算法，使用词袋模型，用 TF – IDF 计算特征的权重，利用 KL 散度计算相似度. 提出了相对相似度机制，避免了引入绝对阈值，适用于多种环境. 实验结果表明，与 3 个流行的 RGB – D SLAM 系统相比，建立的优化模型优化位姿和特征点，提高了定位精度；与利用二进制字典的闭环检测方法相比，本文的闭环检测算法可以准确地检测到闭环，闭环矫正后的精度更高. 实验结果证明了本文方法的有效性.

参考文献：

[1] Reuter S, Dietmayer K, Vo B N, et al. The labeled multi-bernoulli filter[J]. *IEEE Transactions on Signal Processing*, 2014, 62(12): 3246 – 3260.

[2] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. *IEEE Transactions on Robotics*, 2015, 31(5): 1147 – 1163.

[3] Sujan V, Dubowsky S. Efficient information-based visual robotic mapping in unstructured environments[J]. *International Journal of Robotics Research*, 2005, 24(4): 275 – 293.

[4] Janabi-Sharifi F, Marey M. A Kalman-filter-based method for pose estimation in visual servoing[J]. *IEEE Transactions on Robotics*, 2010, 26(5): 939 – 947.

[5] Li S, Ni P. Square-root unscented Kalman filter based simultaneous localization and mapping[C]//IEEE International Conference on Information and Automation. Harbin, 2010: 2384 – 2388.

[6] Lee J S, Nam S Y, Chung G W K. Robust RBPF-SLAM for indoor mobile robots using sonar sensors in non-static environments[J]. *Advanced Robotics*, 2011, 25(9/10): 1227 – 1248.

[7] Triggs B, McLauchlan P F, Hartlry R I, et al. Bundle adjustment a modern synthesis[C]//Vision Algorithms: Theory and Practice. New York, 2000: 298 – 372.

[8] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: real-time single camera SLAM[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 1052 – 1067.

[9] Kim J, Yoon K J, Kim J S, et al. Visual SLAM by single-camera catadioptric stereo[C]//Proceedings of International Joint Conference on SICE-ICASE. Busan, 2006: 2005 – 2009.

[10] Han C, Xiang Z, Liu J, et al. Stereo vision based SLAM in outdoor environments[C]//Proceedings of IEEE International Conference on Robotics and Biomimetics. Sanya, 2007: 1653 – 1658.

[11] Paz L M, Pinies P, Tardos D, et al. Large-scale 6-DOF SLAM with stereo-in-hand[J]. *IEEE Transactions on Robotics*, 2008, 24(5): 946 – 957.

[12] Zhang G, Lee J H, Lim J, et al. Building a 3-D line-based map using stereo SLAM[J]. *IEEE Transactions on Robotics*, 2015, 31(6): 1364 – 1377.

[13] Engel J, Stückler J, Cremers D. Large-scale direct SLAM with stereo cameras[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, 2015: 1935 – 1942.

[14] Henry P, Krainin M, Herbst E, et al. RGB-D mapping: using depth cameras for dense 3D modeling of indoor environments[C]//Proceedings of International Symposium on Experimental Robotics (ISER). Delhi, 2010: 477 – 491.

[15] Concha A, Civer J. RGBDTAM: a cost-effective and accurate RGB-D tracking and mapping system[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, 2017: 6756 – 6763.

[16] Whwlan T, Salas-Moreno R F, Glocker B, et al. ElasticFusion: real-time dense SLAM and light source estimation[J]. *The International Journal of Robotics Research*, 2016, 35(14): 1697 – 1716.

[17] Jafari O H, Mitzel D, Leibe B. Fast ICP-SLAM for a bi-steerable mobile robot in large environments[C]//2015 International Conference on Advanced Robotics. Istanbul, 2015: 5636 – 5643.

[18] Endres F, Hess J, Sturm J, et al. 3-D mapping with an RGB-D camera[J]. *IEEE Transactions on Robotics*, 2014, 30(1): 177 – 187.

[19] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras[J]. *IEEE Transactions on Robotics*, 2017, 33(5): 1255 – 1262.

[20] Gavez-Lopez D, Tardós J D. Bags of binary words for fast place recognition in image sequences[J]. *IEEE Transactions on Robotics*, 2012, 28(5): 1188 – 1197.

[21] Robertso N, Stephe N. Understanding inverse document frequency: on theoretical arguments for IDF[J]. *Journal of Documentation*, 2004, 60(5): 503 – 520.

[22] Chen J N, Matzinger H, Zhai H Y, et al. Centroid estimation based on symmetric KL divergence for multinomial text classification problem[C]//Proceedings of IEEE International Conference on Machine Learning and Applications. Orlando, 2018: 1174 – 1177.

[23] Dong Y, Yao K S, Su H, et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition[C]//IEEE International Conference on Acoustics. Vancouver, 2013: 7893 – 7897.

[24] Anysha V, Meher S. Facial expression recognition using local binary patterns and Kullback-Leibler divergence[C]//Proceedings of IEEE International Conference on Communications and Signal Processing. Melmaruvathur, 2015: 349 – 353.

[25] Toru W, Yamashita Y. Affine-invariant recognition of handwritten characters via accelerated KL divergence minimization[C]//Proceedings of International Conference on Document Analysis & Recognition. Beijing, 2011: 1095 – 1099.

[26] Strasdat H, Montiel J M M, Davison A J. Scale drift-aware large scale monocular SLAM[C]//Proceedings of Robotics: Science and Systems. Zaragoza, 2010: 73 – 80.

[27] Sturm J, Englhard N, Endres F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura, 2012: 573 – 580.