DOI:10.11918/j.issn.0367-6234.201804142

基于图像关键帧的 Visual-Depth Map 建立方法

马 琳, 杨 浩, 谭学治, 冯冠元

(哈尔滨工业大学 电子与信息工程学院, 哈尔滨 150001)

摘 要: 对于室内视觉定位系统,需要在离线阶段建立 Visual Map 数据库用来存储图像信息,在线阶段用户通过与 Visual Map 数据库进行比对来完成用户位置的估计. 离线阶段建立的数据库可以采用逐点采样或视频流采样的方式. 但是无论何种方式,考虑到数据库中图像信息的相似性,传统方式建立的数据库中存储图像有较多冗余,导致增加了在线阶段的定位时间开销. 因此,本文根据 Visual Map 中的相邻图像间的相似性,提出了一种基于图像关键帧的 Visual-Depth Map 建立方法,有效地减少了离线数据库的规模. 在离线阶段,本文使用 Kinect 传感器同时获得图像信息和深度信息;然后,通过基于图像相似度的图像关键帧算法对原始图像序列进行筛选,得到关键帧序列,从而实现 Visual-Depth Map 的建立. 在线阶段,用户可以直接输入查询图像与 Visual-Depth Map 中的图像序列进行检索匹配,找到相似度较高的匹配图像,再通过 EPnP 算法进行 2D-3D 的位姿估计,完成用户位置的计算. 实验证明,本文所提方法可以在保证较高定位精度的前提下,有效减少离线数据库规模,降低在线阶段的定位时间开销.

关键词: 室内定位系统;视觉定位;Visual-Depth Map;关键帧

中图分类号: TP399 **文献标志码**: A **文章编号**: 0367-6234(2018)11-0023-09

Image Keyframe-based Visual-Depth Map Establishing Method

MA Lin, YANG Hao, TAN Xuezhi, FENG Guanyuan

(School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China)

Abstract: In the indoor visual positioning system, the offline Visual Map database is usually used to store the database images, and then user position is estimated by comparing with the images in the Visual Map database in the online phase. The database establishment in the offline phase can be realized with point-by-point sampling method or video stream sampling method. However, taking into account the similarity of database images, redundancy exists between the images in the database, which leads to more positioning time consumption in the online phase. Therefore, owing to the similarity between successive images, a Visual-Depth Map method which reduces the database scale is proposed based on image keyframes. In the offline phase, the method adopts a Kinect sensor which acquires image information and depth information simultaneously. And then, for establishing the Visual-Depth Map, the keyframe sequence is selected from original database images by the image keyframe algorithm that is based on the image similarity. In the online phase, the query image captured by the user is retrieved and matched with the database image in the Visual-Depth Map with higher similarity. The EPnP algorithm is employed to estimate the query camera pose so as to achieve the user position. Experimental results show that the proposed Visual-Depth Map method based on the image keyframe reduces the database scale and positioning time consumption comparing with traditional methods, under the precondition of the high positioning accuracy.

Keywords: indoor positioning system; visual positioning; Visual-Depth Map; keyframe

由于室内视觉定位具有低成本、数据获取便捷、适应性强等特点,近年来引起了国内外学者的广泛关注,并已成为主流定位技术之一. 文献[1]将用户输入图像与图像数据库进行匹配找到最佳匹配图像对,利用对极几何约束进行相机位姿计算,从而得到用户位置,该定位方法达到了米级的定位精度. 文献[2]提出将纠正线与单应矩阵两者相结合应用到室内定位技术中,在保证定位精度的同时,提高了定位速度. 文献[3-4]使用 RGB-D 传感器,在室内场景中建立三维稠密地图,并利用该地图实现室内定位功能,然而,该方法在地图构建时需要使用 GPU 对每一帧图像进行处理才能实现算法实时性. 文献[5]建立了一种利用室内三维数据库以及对极几何算法实现的室内定位系统,该系统在求解用户位置时,没有利用数据库中的深度信息,而仅通过图像信息求解相机位姿. 文献[6]提出一种基于视觉和惯导的定位与建图方法,能够通过在线的闭环检测和

收稿日期: 2018-04-19

基金项目: 国家自然科学基金(61571162)教育部-中国移动科研基金(MCM20170106)黑龙江省自然科学基金(F2016019)

作者简介: 马 琳(1980—),男,副教授,博士生导师;
谭学治(1957—),男,教授,博士生导师

通信作者: 马琳,malin@hit.edu.cn

非线性优化实现相机位姿的全局约束,该方法具有亚米级的定位精度. 文献[6]方法虽然具有较高的定位精度,但该方法除需要视觉传感器之外,还需要其他传感器辅助才能实现用户位置的估计.

综上所述,这些视觉定位方法,无一例外需要图像数据库(Visual Map)的支持. Visual Map 是一种用于室内定位的离线图像数据库,该数据在离线阶段进行创建,该数据库的特点是储存于该数据库的图像同时具有位置信息. 在在线定位阶段,视觉定位方法通过查询图像与 Visual Map 中的图像进行检索与匹配,最终完成用户所在位置的估计. 不难看出,降低查询图像在 Visual Map 数据库的检索时间对于保证视觉定位的实时性具有重要作用[6],而在相同场景下的图像检索时间与 Visual Map 数据库的规模有着十分密切的关系,因此,建立一个规模小、信息完整的 Visual Map 数据库是十分重要的.

目前,Visual Map 的建立方式主要有两种,分别是逐点采样方式和视频流采样方式. 逐点采样方式是通过相机在已经标记的地点进行图像采集并记录对应的位置,文献[1,2,8]中数据库的建立方法便是基于这种方式实现的. 该方法采集数据准确,但是需要耗费巨大的人力物力,而且很难设置一个合理的空间采样间隔,该方法通常会对场景进行大量而密集的图像采集,导致了采集到的图像存在冗余. 视频流采样方式是通过相机对场景进行视频采集,并对视频流采样得到图像序列. 该方法通过计算求出每一帧图像序列所对应的位置信息,从而完成 Visual Map 的建立[9-10]. 文献[11]提出了一种基于视频流的 Visual Map 快速建立方法,该方法通过将视频流进行等间隔采样从而完成离线数据库的建立. 然而,通过视频流采样方式建立 Visual Map 时,通过固定频率采集图像数据时会存在场景信息的冗余或缺失,会影响在线定位时的定位速度和定位精度.

因此,针对上述问题,本文提出一种基于图像关键帧的 Visual-Depth Map 建立方法. 图像关键帧是指 Visual Map 中通过关键帧选择算法所选取的具有代表性的场景图像. 文献[12]提出了一种基于图像关键帧的稠密平面的图像数据库建立方法,但该方法在关键帧选取时只考虑了图像间空间距离信息,忽略了场景内容的变化情况,因而所选取的关键帧存在冗余或缺失的情况. 因此本文提出的图像关键帧算法主要依据图像内容进行关键帧的选取,这样所选取的关键帧能够准确地反映出室内场景的变化情况. 同时,考虑到室内系统的实际需求,本文采用 Kinect 传感器进行信息采集,Kinect 传感器能够同时获得场景中的图像信息和深度信息. 因此,本文同时使用图像信息与深度信息进行 Visual-Depth Map 的创建. 在在线阶段,将用户查询图像与 Visual-Depth Map 中图像信息和深度信息进行匹配,运用 EPnP 算法[12]进行用户位置的估计. 实验结果表明,在室内定位中使用本文提出的方法进行数据库建立时,减少了数据库中的冗余信息,并缩短了在线定位时间,同时保证了用户的定位精度.

1　室内视觉定位系统

1.1　系统框架

本文提出的室内视觉定位系统主要分为离线和在线两个阶段,见图1. 在离线阶段,通过图像关键帧算法对 Kinect 传感器采集的图像和深度信息进行选取从而完成 Visual-Depth Map 的建立工作;在在线阶段,用户输入的查询图像通过特征点提取、检索匹配和 EPnP 算法求解相机位姿后可以得到该用户所在的位置信息.

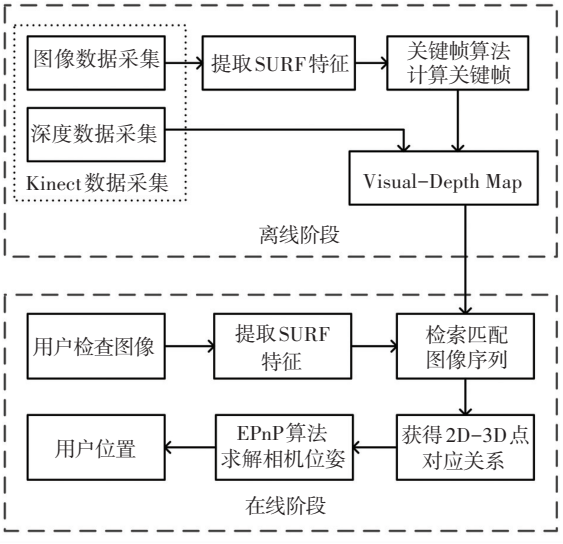


图1　室内视觉定位系统框图

Fig.1　Framework of indoor visual localization system

与传统的离线阶段使用摄像头采集图像信息的方法不同,本文提出的建立视觉定位数据库采用的传感器是 Kinect,它能同时获取图像信息和深度信息,因此在线定位时的场景信息更加丰富,可以获得较好的定位结果. 为了减少数据库的规模,在离线阶段利用图像关键帧算法在 Kinect 采集得到的原始图像序列中选取关键帧序列,并只将关键帧序列所对应的信息存储到数据库中,从而减小数据库的规模. 与此同时,虽然在离线阶段建立了 Visual-Depth Map,存储了关键帧中特征点的图像特征信息与深度信息,但是在在线阶段,用户同样使用一张查询图像,就可以完成用户位置的计算工作. 在计算

用户位置时，不同于通过对极几何算法求解，而是使用 EPnP 算法进行求解，该算法能够利用空间 3D 和 2D 特征点进行位姿的求解，算法复杂度较低，能够较快地完成位姿的求解，从而得到用户的位置.

1.2　离线阶段

在离线阶段，主要完成 Visual-Depth Map 视觉定位数据库的建立工作.

首先，通过 Kinect 传感器采集场景的原始图像信息和深度信息，通过图像关键帧算法计算每一个位置点所对应的图像信息能否作为图像关键帧存储到 Visual-Depth Map 中，而后将满足条件的图像关键帧中的特征点提取出来，并找到特征点所对应的深度信息，并将这些信息与位置信息存入 Visual-Depth Map 数据库中. 原始图像序列逐一通过图像关键帧算法筛选后，便完成了 Visual-Depth Map 的建立工作.

通过以上方法建立的 Visual-Depth Map 包含了场景中的关键帧所对应的位置信息、关键帧的特征点信息以及特征点所对应的深度信息，这些信息将会在在线阶段用于用户在室内场景中的位置计算.

1.3　在线阶段

在在线阶段，主要完成根据用户查询图像进行用户位置计算的工作.

首先，利用 Visual-Depth Map 中的图像信息，将用户查询图像进行特征点提取，并将其与数据库中图像关键帧的特征点进行匹配，得到与查询图像最相匹配的关键帧. 同时，记录用户查询图像的 2D 特征点信息和与之相匹配的图像关键帧的 2D 特征点信息，并查询得到其所对应的空间 3D 点信息，从而得到用户输入图像的 2D 特征点与数据库中 3D 点的匹配. 最后，利用数据库中 3D 点与用户查询图像中的 2D 特征点的对应关系，通过 EPnP 算法计算得到用户的位置信息.

2　图像关键帧算法建立 Visual-Depth Map 与定位

2.1　离线阶段传统数据库建立方法

传统 Visual Map 建立通常采用逐点采样或视频流采样方式实现. 这两种建立方法所得到的 Visual Map 中都会存储带有位置标签的图像序列. 假设，所建立的 Visual Map 中共存储 n 组数据，那么便可以从 Visual Map 中得到采集的图像序列 $\{I_1, I_2, I_3, \cdots, I_n\}$，其中 I_i 表示在图像序列中编号为 i 的图像信息，同时 Visual Map 中也存储着图像信息所对应的位置信息序列 $\{l_1, l_2, l_3, \cdots, l_n\}$，其中 l_i 表示编号为 i 的图像信息所对应的位置信息. 与此同时，为了使 Visual Map 的图像在在线阶段能够与用户查询图像进行检索匹配，在离线阶段还需对每幅图像中的图像特征点提取，并将特征点信息用特征向量表示.

为了避免采集的图像信息数据量过于庞大，往往在离线阶段数据库建立时，采集的图像位置具有一定的空间间隔. 通过逐点采样建立 Visual Map 时，通常相邻的采样间隔满足式(1)：

$$d = \| l_i - l_{i-1} \|_2. \tag{1}$$

式中 d 为预先设定的采样距离. 通过式(1) 的采样条件，可以进行等间隔的数据采样，采样的位置在空间上均匀分布.

另外，通过视频流采样的方式建立 Visual Map 时，由于很难直接得到相机在移动过程中图像所对应的位置信息，往往使采集设备匀速前进，从而实现视频流的采集，最后进行图像帧的提取和图像帧所对应的位置坐标计算为

$$\begin{cases} X_n = X_0 + v\dfrac{n}{m}\cos\theta, \\ Y_n = Y_0 + v\dfrac{n}{m}\sin\theta. \end{cases} \tag{2}$$

式中：m 表示采集的视频的帧速率，v 为数据采集平台运动速度，θ 为数据采集平台运动方向和坐标系 X 轴夹角，(X_0, Y_0) 表示建立参考坐标系的坐标原点，(X_n, Y_n) 即为第 n 帧图像所对应的位置坐标. 通过式(2) 可知，当已知匀速运行速度 v、帧速率 m 和设备运行方向时，通过帧序号 n 的查找便可以计算得到该图像帧所对应的位置信息.

但是，以上两种 Visual Map 建立方法，在参数选择方面往往没有规范化的要求，也没有考虑场景中内容的多样性和复杂性，因此以上两种方法所选取的数据库图像可能无法全面准确地反映出场景内容信息，影响定位精度；也可能存在采样点采集过于密集的情况，造成数据库规模过大，影响定位速度.

2.2　离线阶段结合图像内容的关键帧算法

由于传统 Visual Map 建立方法在定位精度和定位速度上存在较大的不确定性，因此本文提出一种在离线阶段结合图像内容的关键帧算法，选取原始图像中的图像关键帧，并引入图像中特征点的深度信息，从而完成 Visual-Depth Map 的建立工作. 在 Visual-Depth Map 中，每组数据中包含图像关键帧所对应的位置信息和图像关键帧中特征点的特征向量以及深度信息. 同时，通过图像关键帧算法所选择的图像关键帧所对应的序号称为关键帧序号 f_k ($k = 1, \cdots, m$，其中 m 为数据库中关键帧总数).

图像关键帧算法首先需要对图像的内容信息进行

表示. 本方法采用 Speeded Up Robust Feature(SURF)特征[14]对图像信息进行描述,其具有特征点识别率高、运行速度快以及对视角、光照、尺度具有较高鲁棒性等优点. 通过 SURF 特征提取算法对图像进行特征点提取,每幅图像中 SURF 特征点对应的特征向量用 s 表示.

当完成图像中 SURF 特征提取后,便可将一幅图像中的信息用若干个 SURF 特征点表示,这样在进行两幅图像中内容比较时,便可以通过 SURF 特征点匹配来判断场景是否存在相似之处. 图像序号为 a 和 b 的两幅图像 I_a 和 I_b 中的 SURF 特征点间的距离为

$$Ed_{ij} = \| s_i - s_j \|_2 \quad s_i \in I_a, s_j \in I_b. \tag{3}$$

当完成 SURF 特征点距离的计算后,便可以通过两点为匹配点对的判决条件式(4)进行特征点是否匹配的判断[14]:

$$\frac{Ed_{\min 1}}{Ed_{\min 2}} \leqslant T_{Ed}. \tag{4}$$

式中:$Ed_{\min 1}$ 为最近邻特征点距离;$Ed_{\min 2}$ 为次近邻特征点欧式距离;T_{Ed} 为正确匹配判决阈值.

完成两幅图像的特征点匹配工作后,便可计算出两幅图像 I_a 和 I_b 之间的匹配点数量,记为 $N_{(a,b)}$. 当两幅图像进行特征点匹配时,若匹配点数量 $N_{(a,b)}$ 占图像中总特征点数越多,两幅图像的相似程度越高,因此图像间匹配点数量 $N_{(a,b)}$ 对于认知两幅图像的相似程度起着十分重要的作用. 本文利用该信息来描述图像间的相似程度,提出图像内容相似度的概念,定义如式(5)所示:

$$S(I_a, I_b) = \frac{2N_{(a,b)}}{N_a + N_b}. \tag{5}$$

式中:N_a 和 N_b 分别表示图像 I_a,I_b 中提取的 SURF 特征点的个数,$N_{(a,b)}$ 为两幅图像匹配特征点个数.

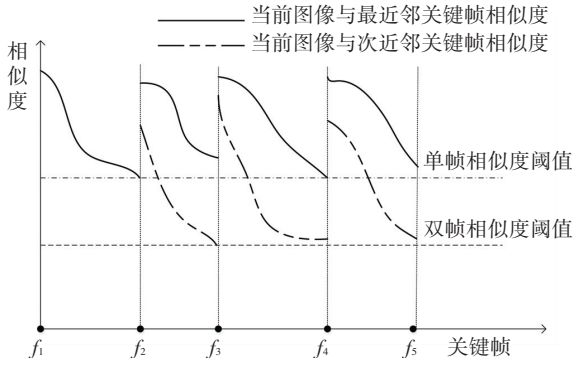


图 2　图像关键帧算法原理

Fig.2　Keyframe algorithm schematic

已知两张图像的相似度后,便可以根据相似度指标从原始图像序列中选取关键帧. 本文所提出的图像关键帧算法原理见图 2. 其主要思想是通过当前图像与最近邻和次近邻关键帧的图像相似度来判断原始图像是否选取为图像关键帧,选取依据是原始图像与最近邻图像关键帧相似度小于一定数值或其与次近邻图像关键帧相似度小于一定数值,以保证图像关键帧间图像相似度差异小,并且尽可能减少数据库中场景信息的冗余和缺失.

提出的图像关键帧算法满足条件见式(6):

$$f_k = \begin{cases} 1 \quad k = 1, i = 1; \\ \underset{i}{\operatorname{argmin}}(S(I_i, I_{f_1}) \leqslant \alpha) \quad k = 2, i \geqslant 2; \\ \underset{i}{\operatorname{argmin}}(S(I_i, I_{f_{i-1}}) \leqslant \alpha, S(I_i, I_{f_{i-2}}) \leqslant \beta)\ k > 2, i \geqslant f_2. \end{cases} \tag{6}$$

式中:i 为原始图像序列序号,k 为关键帧序号,α 为单帧阈值参数,β 为双帧阈值参数. 单帧阈值是指关键帧选取准则中原始图像与最近邻图像关键帧相似度值的最高门限,双帧阈值是指关键帧选取准则中原始图像与次近邻图像关键帧相似度值的最高门限.

由图像关键帧算法原理可知,原始图像与最近邻和次近邻图像关键帧的相似度随着它们之间距离的增大而不断减小,当减小到一定程度时,如果满足原始图像与最近邻图像关键帧相似度小于单帧阈值 α 或其与次近邻图像关键帧相似度小于双帧阈值 β 时,那么当前的原始图像便可以选取为图像关键帧,其所对应的图像序号便可以选取为关键帧序号. 由此可知,单帧阈值 α 和双帧阈值 β 对于关键帧选取非常重要,它们所设置的大小直接决定数据库的规模,后文将对这两个参数的选取进行分析.

由式(6)计算得到 Visual-Depth Map 建立所需要的关键帧序列 $\{f_1, f_2, f_3, \cdots, f_m\}$,该序列表示了 Visual-Depth Map 中所需要存储的图像关键帧所对应的序号. 通过图像关键帧序列便可以将图像关键帧所对应的位置信息、图像关键帧中 2D 特征点和对应的空间 3D 点存储到数据库中,从而完成基于图像关键帧的 Visual-Depth Map 数据库建立工作. 基于此方法建立的 Visual-Depth Map 内容见表 1.

表 1　基于图像关键帧方法建立的 Visual-Depth Map

Tab.1　Visual-Depth Map established by image keyframes

关键帧序号	地理位置	图像 2D 特征点	空间 3D 点
f_1	l_{f_1}	s_{11}、s_{12}、$s_{13}\cdots$	p_{11}、p_{12}、$p_{13}\cdots$
f_2	l_{f_2}	s_{21}、s_{22}、$s_2\cdots$	p_{21}、p_{22}、$p_{23}\cdots$
……	……	……	……
f_m	l_{f_m}	s_{m1}、s_{m2}、$s_{m3}\cdots$	p_{m1}、p_{m2}、$p_{m3}\cdots$

2.3　基于 EPnP 的在线定位算法

在在线定位阶段,用户输入查询图像后,将查询

图像中提取的 SURF 特征点与 Visual-Depth Map 数据库中的不同图像关键帧的特征点进行匹配，得到与查询图像最匹配的前 N 张图像关键帧. 数据库与用户图像的匹配特征点间是平面点与平面点之间的对应关系，即 2D－2D 的对应关系，再利用 Visual-Depth Map 中的空间 3D 点，找到数据库中的 2D 点对应的 3D 点，从而实现了用户查询图像中的 2D 点与 Visual-Depth Map 中的 3D 点间的匹配. 当得到 2D 点与 3D 点的匹配点对后，数据库图像与用户查询图像的相机之间的位姿关系便可以通过 EPnP 算法进行求解. EPnP 算法的核心思想是将所有点分解成四个控制点加权和的形式，通过小孔成像模型，建立起 3D 点和 2D 点之间的关系式，多个特征点联合进行求解，得到位姿关系. 这里的位姿关系主要是指旋转关系和平移关系，分别用旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$ 来表示.

根据小孔成像模型，可知参考点在相机坐标系到像素坐标系的投影关系如式(7)所示

$$w\begin{bmatrix} u^c \\ 1 \end{bmatrix} = \boldsymbol{A}p^c. \tag{7}$$

式中：w 为缩放因子，$\boldsymbol{A}$ 为相机内参矩阵，在世界坐标系下的 3D 点的坐标和与之对应的在相机坐标系下的 2D 点坐标分别由 p^c 和 u^c.

利用 EPnP 算法，假设所有匹配点在世界坐标系由 4 个控制点表示为

$$p^w = \sum_{i=1}^{4} a_i c_i^w. \tag{8}$$

同理，其匹配点在用户相机坐标系应满足：

$$p^c = \sum_{i=1}^{4} a_i c_i^c. \tag{9}$$

式中 a_i 为控制点权重值，满足：

$$\sum_{i=1}^{4} a_i = 1\ . \tag{10}$$

因此，可将式(7)中的世界坐标系下的 3D 点 p^c 用式(9)表示，并将相机坐标系下的 3D 点的坐标为 $[x_j^c, y_j^c, z_j^c]^{\mathrm{T}}$，在像素坐标系下的 2D 点的坐标为 $[u, v]^{\mathrm{T}}$，f_u、f_v 为相机的焦距参数，u_c、v_c 为图像中像素的平移参数，展开如式(11)所示：

$$w\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & u_c \\ 0 & f_v & v_c \\ 0 & 0 & 1 \end{bmatrix} \sum_{i=1}^{4} a_i \begin{bmatrix} x_i^c \\ y_i^c \\ z_i^c \end{bmatrix} \tag{11}$$

由式(11)最后一行展开可知：

$$w = \sum_{i=1}^{4} a_i z_i^c. \tag{12}$$

由式(12)的约束条件可以将一个控制点的约束式(11)表示为：

$$\begin{cases} \sum_{i=1}^{4} a_i [f_u x_i^c + a_i (u_c - u) z_i^c] = 0, \\ \sum_{i=1}^{4} a_i [f_v y_i^c + a_i (v_c - v) z_i^c] = 0\ . \end{cases} \tag{13}$$

因此，利用式(13)可以求出用户相机坐标系下 2D 点对应的 3D 点的位置[13]，即完成了用户相机坐标系下的 2D 点 u^c 到用户相机坐标系下的 3D 点 p^c 之间的坐标转换.

当求出与 p^w 对应的 p^c 时，可通过迭代最近点(Iterative Closest Point，ICP)算法[15]进行旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{t}$ 的计算，当得到 $\boldsymbol{R}$ 和 $\boldsymbol{t}$ 时，便得到用户输入查询图像的相机坐标系和 Visual-Depth Map 中图像的相机坐标系之间的位姿关系.

本文采用的定位算法需要找到和用户查询图像最相匹配的前 4 张图像，由于已知 Visual-Depth Map 中图像的位置信息，因此通过位置信息和平移向量 t 确定经过用户查询图像的相机坐标与 Visual-Depth Map 中图像所对应的相机坐标的 4 条直线，将 4 条直线统一到世界坐标系下可表示为

$$a_i x + b_i y + c_i = 0.$$

式中 i 表示第 i 条直线，$i = 1, 2, 3, 4$.

当计算出经过用户坐标与匹配图像所对应相机坐标的直线后，用户位置$(x_{\mathrm{user}}, y_{\mathrm{user}})$便可通过计算距离直线最近坐标点的方法完成计算，计算公式为

$$(x_{\mathrm{user}},\ y_{\mathrm{user}}) = \min_{x,\,y} \sum_{i=1}^{4} d_i(x,y)\ . \tag{15}$$

式(15)中，

$$d_i(x,y) = \frac{|a_i x + b_i y + c_i|}{\sqrt{a_i^2 + b_i^2}}\ . \tag{16}$$

3　实验仿真

为验证本文所提出算法的可靠性与适用性，选择哈尔滨工业大学通信技术研究所 12 楼作为实验场所. 实验场所的平面见图 3.

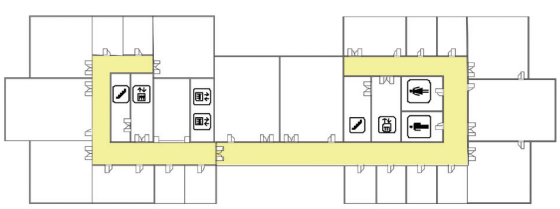


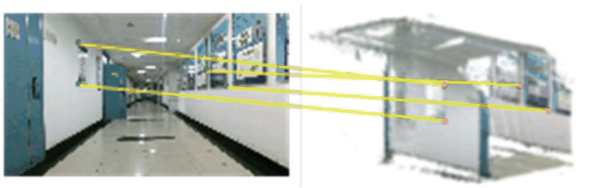
图 3　实验场所的平面

Fig.3　Floor plan of the experimental area

在实验场景中，通过 Kinect 传感器进行图像信息和深度信息的采集，并同时记录采集数据时所对

应的位置信息,本文每隔 0.5 m 对场景的图像和深度信息进行采集. Kinect 传感器能够同时采集场景中的图像信息与深度信息,并得到图像 2D 特征点与对应的空间 3D 点之间的对应关系,从而实现 Visual-Depth Map 中所需的图像 2D 特征点信息深度信息(即空间点的 3D 位置)的获取,为在线阶段的 EPnP 定位算法提供了数据基础. Kinect 进行场景数据采集时,图像特征点与对应的 3D 点见图 4. 图 4 中包含一副实验场景的图像以及一副含有深度信息的三维点云图像,两者为同一时刻采集得到的. 通过图 4 可以看出,从图像中提取的 SURF 特征点与点云中的 3D 点一一对应,即可以通过点云图像直接得到 SURF 特征点的空间位置.

利用图像特征点信息可以进行基于图像关键帧算法的原始图像筛选工作,进而通过图像特征点和对应 3D 点可实现位置估计. 因此,通过 Kinect 采集得到的原始数据可以满足场景信息分析、Visual-Depth Map 建立以及用户位置计算的需求.



(a) 场景图像　　(b) 场景点云

图 4　图像特征点与对应的 3D 点示意

Fig.4　Diagram of image feature points and the corresponding 3D points

通过 Kinect 完成实验场景的数据采集工作后,需要对基于图像关键帧算法的必要性进行分析. 在 Kinect 传感器所采集得到的原始图像序列中,选取出两组相邻图像对进行相似度分析,所选相邻图像对见图 5. 相邻图像对均为相隔距离为 0.5m 时拍摄得到的,在此分别计算两组(A 组、B 组)相邻图像对中的特征点数量、匹配特征的数量、图像相似度等信息,具体实验数据见表 2.

通过表 2 对比发现,在实际场景图像序列中,纹理特征较多的图像中能提取出的特征点数量较多,反之亦然. 在不同地点进行等间隔采样时,相邻图像对 A 间的相似度为 0.375,相邻图相对 B 间的相似度为 0.441,可以看出等间隔采样的图间的相似度具有一定的差异.

正是由于这种图像内容的差异性,等间隔采样建立数据库时很难准确地描述场景信息,或者容易造成冗余信息过多的情况. 因此,本文提出一种基于图像关键帧的视觉定位数据库建立方法,其利用图像关键帧对场景信息进行表示,并根据图像间的相似程度来选取数据库中的关键帧. 该方法在保证定位精度的前提下,减小了数据库的规模,从而降低了在线定位的时间开销.

(a) 相邻图像对 A

(b) 相邻图像对 B

图 5　两组相邻图像对

Fig.5　Two groups of successive image pairs

表 2　相邻图像对相似度分析

Tab.2　The similarity analysis of successive image pairs

相邻图像对	图像序号	匹配特征点数	图像特征点数	图像相似度
相邻图像对 A	22 23	927 976	357	0.375
相邻图像对 B	37 38	812 954	389	0.441

在本文提出的图像关键帧算法中,阈值参数的选择至关重要. 因此本文对关键帧算法中的单、双帧阈值 α、β 的选取进行了仿真实验. 通过计算相邻图像相似度发现,图像相似度在 0.4 左右波动,因此在实验中相似度值 0.4 为 α 的上界,并且 α 分别取 0.4、0.35、0.3、0.25, 由于 β 的值应小于 α,因此 β 分别取 $\alpha - 0.05$、$\alpha - 0.10$、$\alpha - 0.15$、$\alpha - 0.20$. 对选定的每组阈值进行关键帧算法的计算,从而得出选取不同阈值时所对应的图像关键帧数量,并不同单、双帧阈值时的相邻关键帧相似度的均值和方差进行计算分析,具体实验数据见表 3、4.

表 3　不同单、双帧阈值时相邻关键帧相似度均值

Tab.3　The similarity mean of successive keyframes choosing different single and double frame thresholds

单帧域值	双帧域值			
	$\beta = \alpha - 0.05$	$\beta = \alpha - 0.10$	$\beta = \alpha - 0.15$	$\beta = \alpha - 0.20$
$\alpha = 0.40$	0.376	0.361	0.358	0.347
$\alpha = 0.35$	0.361	0.335	0.300	0.297
$\alpha = 0.30$	0.327	0.286	0.265	0.261
$\alpha = 0.25$	0.284	0.240	0.224	0.224

表 4　不同单、双帧阈值时相邻关键帧相似度方差

Tab.4　The similarity variance ofsuccessive keyframes choosing different single and double frame thresholds

单帧域值	双帧域值			
	$\beta = \alpha - 0.05$	$\beta = \alpha - 0.10$	$\beta = \alpha - 0.15$	$\beta = \alpha - 0.20$
$\alpha = 0.40$	0.001 25	0.001 55	0.001 46	0.001 46
$\alpha = 0.35$	0.001 55	0.002 94	0.001 50	0.001 15
$\alpha = 0.30$	0.002 47	0.002 57	0.000 92	0.000 70
$\alpha = 0.25$	0.003 19	0.001 34	0.001 02	0.001 02

可以看出，当单帧阈值不变，减少双帧阈值时，相邻关键帧相似度均值减小，相邻关键帧相似度方差也在减少，说明在这种情况下，可以减少数据库相邻帧的相似度差异，并且保证数据的波动性较小；当单、双阈值的差保持不变时，同时减小两者数值，也会造成相邻关键帧相似度均值的减小，但是会影响数据的波动性.

本文还对不同单、双帧阈值时选取的关键帧数量进行了比较，见图 6. 由实验数据可知，当选取不同的阈值时，所得到的关键帧的数量、相邻关键帧相似度的均值和方差均会受到影响. 而本文所希望得到的图像关键帧，应在满足约束条件下，符合相邻关键帧相似度波动性较小并且关键帧数量较少的要求，换言之，本文所需得到的关键帧，其可通过较少的图像数量来更加平稳地描述实验场景的图像信息.

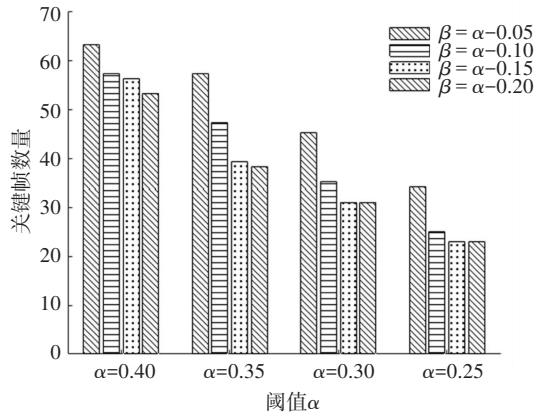


图 6　不同单、双帧阈值时选取的关键帧数量

Fig.6　The number of selected keyframes choosing different single and double frame thresholds

因此，综合考虑关键帧的数量、相邻关键帧相似度的均值和方差，本文选取 $\alpha = 0.35$、$\beta = 0.2$ 作为阈值参数，该组阈值不仅选取出了较少数量的关键帧，也保证了相邻关键帧的相似度较高，并且数据的波动性较小. 因此，该组阈值参数适用于图像关键帧算法的计算.

经实验计算，在建立数据库时，采用逐点采样方法和基于图像关键帧方法所选取的图像序号见表 5. 可看出，采用逐点采样方法，每隔 0.5 m 采集一组数据，在前 10 m 共采集 20 张图像，而经过基于关键帧方法筛选后，前 20 张图像中选取 11 张作为关键帧来反映场景信息. 因此，通过该方法所建立的数据库规模更小. 在该算法中，关键帧的选取并不是均匀的，而是一种基于图像内容的非均匀选择方式.

表 5　不同数据库建立方法所选取的图像序号

Tab.5　The selected image sequence number of the different database selection method

数据库建立方法	所选取的图像序号
逐点采样方法	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
基于关键帧方法	1,3,4,6,8,10,12,13,15,17,19

在实验场景中，对基于图像关键帧的 Visual-Depth Map 建立方法进行了数据库规模分析. 在此将逐点采样方式、视频流方式以及本文提出的基于关键帧算法的数据库建立方式进行比较，结果见表 6. 当采用逐点采样方式时，每隔 0.5m 采集一组数据，一共采集了 65 组数据，而经过关键帧算法进行筛选以后，数据库中的数据大小仅为 34 组，数据库规模减少了 47.7%. 当采用视频流方式时，可以通过设置采样频率，在相同实验场景中采集相近数量的数据，本实验中一共采集了 64 组数据，经关键帧算法筛选后，数据库中的数据容量仅为 33 组，数据库规模减少了 48.4%. 在在线阶段，定位时间的长短主要取决于用户查询图像检索匹配所消耗的时间，而检索匹配时间主要取决于视觉定位数据库的规模. 因此，当数据库规模明显减小时，用户查询图像在线阶段用来进行检索匹配的时间也会随之减少，从而降低了在线定位的时间开销.

表 6　不同数据库建立方法数据库规模对比

Tab.6　Comparison of database scale by different establishing methods

数据库建立方法	数据库中数据量/组
逐点采样	65
逐点采样+关键帧	34
视频流	64
视频流+关键帧	33

另一方面，本文对不同定位算法的定位结果进行了比较和分析. 用户输入的查询图像见图 7，在所建立的平面直角坐标系 XOY 中，该用户所在的位置 l 为(7 m, 1 m)，即用户真实位置. 通过不同定位方法得到的定位结果见图 8，其中基于对极几何算法计算得到的用户位置坐标 l_1 为(6.75 m, −0.44 m)，基于 EPnP 算法计算得到的用户位置坐标 l_2 为

(6.98 m, 0.93 m). 相较于对极几何算法,同样使用一张查询图像,基于 EPnP 的在线定位算法可以获得较高的定位精度.

图 7 用户查询图像

Fig.7 User query image

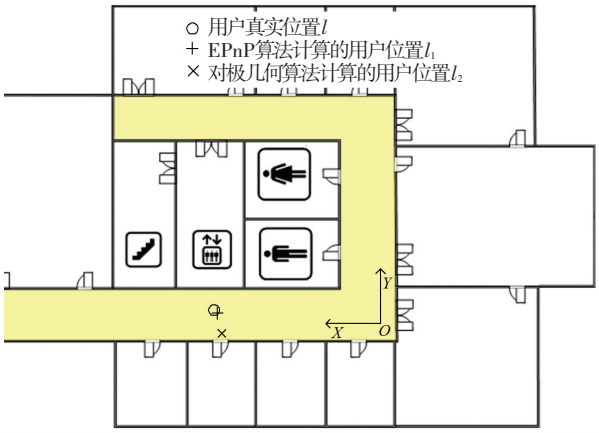


图 8 定位结果

Fig.8 Diagram of localization results

最后,在实验场景中对本文提出的基于图像关键帧的 Visual-Depth Map 建立方法进行定位误差累计概率的计算,其对比结果见表 7,误差累计概率曲线见图 9. 本方法定位结果的误差小于 1 m 的概率为 47%,小于 2 m 的概率为 84%.

图 9 分别对离线阶段使用逐点采样、视频流采样和关键帧方法所建立的数据库进行了比较,同时在在线阶段通过 EPnP 或对极几何算法进行定位. 通过对比可以发现,离线数据库中加入深度信息形成 Visual-Depth Map 后,使用 EPnP 算法的定位精度明显高于对极几何方法的定位精度. 与此同时,通过关键帧算法选择后所建立的数据库与逐点采样和视频流采样方法发现,其定位精度基本不变,表 7 中详细比较了不同数据库建立方法的定位精度,可以看出,关键帧方法所建立的数据库的平均定位误差与不使用关键帧方法时基本相同. 因此,采样关键帧算法建立 Visual-Depth Map 时,在减少数据库规模的情况下,并没有过多的降低定位精度,该结果说明所选取图像关键帧减少了数据库中冗余信息.

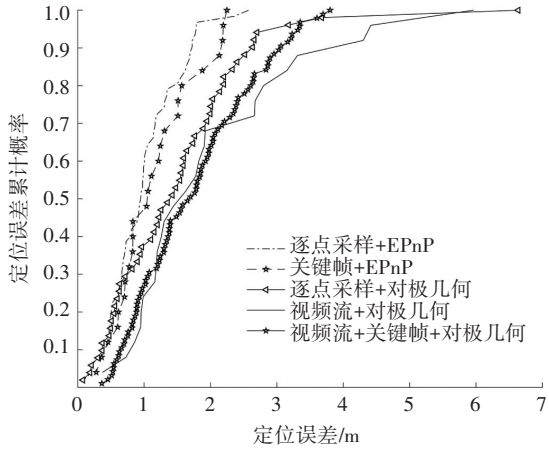


图 9 定位误差累计概率

Fig.9 CDFs of localization errors

表 7 不同数据库建立方法定位精度对比

Tab.7 Comparison of online positioning accuracy for different databases establishing methods

数据库建立方法	1 m 以内	2 m 以内	3 m 以内	平均误差/m
逐点采样+Visual Map	37%	72%	94%	1.71
视频流+Visual Map	24%	68%	80%	1.75
视频流+关键帧+Visual Map	25%	61%	85%	1.76
逐点采样+Visual-Depth Map	60%	97%	100%	0.99
关键帧+Visual-Depth Map	47%	84%	100%	1.17

综上所述,通过 Visual-Depth Map 建立数据库时,由于加入了图像的深度信息,并用 EPnP 算法进行位姿计算,定位精度有了明显地提高. 与此同时,通过比较逐点采样方法和关键帧算法进行 EPnP 定位时,可以发现,使用关键帧算法的定位结果与采用逐点采样方法的定位结果基本相当,而且它的定位精度高于采用逐点采样并用对极几何算法的定位方法. 以上分析说明图像关键帧算法所选择的关键帧序列能够反映场景内容信息,具有一定代表性.

4 结 论

本文提出一种基于图像关键帧的 Visual-Depth Map 建立方法. 本文所提出的图像关键帧算法主要依据图像内容进行关键帧的选取,所选取的关键帧能够准确地反映出室内场景的变化情况. 该方法实现的室内定位系统相较于基于对极几何算法的室内

定位系统，定位精度明显提高，定位误差小于 1 m 的概率为 47%，小于 2 m 的概率为 84%. 与此同时，本文利用图像关键帧算法实现了原始图像的筛选，在保证定位精度的同时，减小了 Visual-Depth Map 数据库的规模，从而降低了在线定位的时间开销.

参考文献

[1] SADEGHI H, VALAEE S, SHIRANI S. A weighted KNNepipolar geometry-based approach for vision-based indoor localization using smartphone cameras[C]// IEEE 8th Sensor Array and Multichannel Signal Processing Workshop. Coruna: IEEE Computer Society, 2014: 37. DOI: 10.1109/SAM.2 014.6882332

[2] GUAN Kai, MA Lin, TAN Xuezhi, et al. Vision-based indoor localization approach based on SURF and landmark [C] //2016 International Wireless Communications and Mobile Computing Conference, 2016 IWCMC.Paphos: Institute of Electrical and Electronics Engineers Inc., 2016: 655. DOI: 10.1109/IWCMC. 2016.7577134

[3] WHELAN T, KAESS M, JOHANNSON H, et al. Real-time large-scale dense RGB-D SLAM with volumetric fusion[J]. International Journal of Robotics Research, 2015, 34(4/5):598. DOI:10.1177/0278364914551008

[4] THOMAS W, STEFAN L, RENATO S, et al.ElasticFusion: dense SLAM without a pose graph[C]//Conference on Robotics: Science and Systems. Rome, Italy. 2015.DOI: 10.15607/rss.2015.xi.001

[5] FENG G, MA L, TAN X. Visual map construction using RGB-D sensors for image-based localization in indoor environments [J]. Journal of Sensors, 2017, 2017 (99): 1. DOI: 10.1155/2017/8037607

[6] ANTON K, FRANCIS E, JORG S, et al. Keyframe-based visual-inertial online SLAM withrelocalization[J]. 2017 International Conference on Intelligent Robots and Systems,2017 IROS. 2017. DOI: 10.1109/IROS.2017.8206581

[7] ANA C, GAUTAM S, JANA. K, et al. Localization in urban environments using a panoramic gist descriptor[J]. IEEE Transactions on Robotics, 2013, 29 (1): 146. DOI: 10.1109/TRO.2012.2220211

[8] HUITL R, SCHROTH G, HILSENBECK S, et al. TUM indoor: An extensive image and point cloud dataset for visual indoor localization and mapping[C]// IEEE International Conference on Image Processing. 2013 ICIP, Melbourne, Australia. 2013:1773. DOI: 10.1109/ICIP.2012.6467224

[9] KIN L, PAUL N. Loop closure detection in SLAM by combining visual and spatial appearance[J]. Robotics & Autonomous Systems, 2006, 54(9):740. DOI: 10.1016/j.robot2006.04.016

[10] PAUL N, KIN L. SLAM-Loop Closing with Visually Salient Features[C]// 2005 IEEE International Conference on Robotics and Automation, 2005 ICRA. Barcelona, Spain. IEEE, 2005: 635. DOI:10.1109/robot.2005.1570189

[11] XUE Hao, MA Lin, TAN Xuezhi. A fast visual map building method using video stream for visual-based indoor localization [C] // 2016 International Wireless Communications and Mobile Computing Conference, 2016 IWCMC.Paphos: Institute of Electrical and Electronics Engineers Inc., 2016: 650. DOI:10.1109/IWCMC .2016.7577133

[12] MING H, WESTMAN E, ZHANG G, et al. Keyframe-based dense planar SLAM[C]// 2017 IEEE International Conference on Robotics and Automation. 2017 ICRA, Singapore, 2017:5110. DOI: 10.1109/icra.2017.7989597

[13] FRANCESC M, VINCENT L, PASCAL F. Accurate non-iterative O (n) solution to the PnP problem[C] // IEEE, International Conference on Computer Vision, 2007 ICCV. Rio de Janeiro, Brazil. IEEE, 2007:1. DOI: 10.1109/iccv.2007. 4409116

[14] BAY H, ESS A, TUYLARRS T, et al. Speeded-up robust features (SURF) [J]. Computer Vision and Image Understanding, 2008, 110(3): 404. DOI: 10.1016/j.cviu.2007.09.014

[15] SZYMON R, MARC L. Efficient variant of the ICP algorithm[J]. Third International Conference on Digital Image Processing, 2001: 145. DOI:10.1109/IM.2001.924423

（编辑　苗秀芝）