移动机器人闭环检测的视觉字典树金字塔 TF-IDF 得分匹配方法

李 博[1] 杨 丹[2] 邓 林[3]

摘 要 针对移动机器人视觉闭环检测中, 基于视觉字典本的场景外观表征性能受制于有限单词个数以及算法效率低的不足, 本文对机器人视觉特征分层量化, 构建视觉字典树, 计算树节点的 TF-IDF 熵作为对应视觉单词的权重, 生成图像 – 单词逆向文档索引. 为消除视觉字典本的单尺度量化误差, 并克服基于字典树投影路径的平面匹配模式中不区分不同层次节点的区分度对闭环检测的影响, 本文融合字典树低层单词的强表征性和高层单词的强鲁棒性, 提出由下而上逐层计算图像间相似性增量的金字塔得分匹配方法. 将不同时刻相似性大于阈值的图像位置提取为候选闭环, 通过后验确认操作剔除误正闭环. 在移动机器人视觉闭环检测实验中, 本文算法提高了图像相似性计算的效率和准确性, 提高了闭环检测的准确率和召回率.

关键词 闭环检测, 视觉字典树, TF-IDF 得分准则, 金字塔匹配

DOI 10.3724/SP.J.1004.2011.00665

Visual Vocabulary Tree with Pyramid TF-IDF Scoring Match Scheme for Loop Closure Detection

LI Bo[1] YANG Dan[2] DENG Lin[3]

Abstract The performance of visual environment modeling in appearance-based robot loop closure detection by using conventional vocabulary is restricted by limited number of visual words and high computational cost. We construct a visual vocabulary tree by clustering the visual features hierarchically captured by a mobile robot. The TF-IDF entropy for each node is computed and is treated as the weight of each visual word, and the inverted index of image-word is exploited. To avoid the quantization error of single scale vocabulary and the neglect of the different discriminative power among different level words of tree-based match, we take advantage of the robustness of high level words and the discriminability of low level words to present a pyramid scoring match scheme. The candidates of loop closures are detected by using a similarity threshold. A posteriori management helps discard outliers by verifying that the two images of the loop closure satisfy some hypothesis constraints. The experiments of loop closure detection in mobile robotics demonstrate that our scheme improves similarity calculation significantly in both accuracy and efficiency and obtains a higher precision-recall ratio with a faster speed of loop closure detection compared to the traditional methods.

Key words Loop closure detection, visual vocabulary tree, TF-IDF scoring scheme, pyramid match

移动机器人在未知环境中根据自身位置和传感器数据创建环境地图, 指导机器人自主定位和导航, 也即机器人同步定位与地图构建 (Simultaneous localization and mapping, SLAM), 是实现真正自主移动机器人的关键. 闭环 (Loop closure) 检测是 SLAM 的基础问题之一, 如何准确判断机器人当前位置是否位于已经访问过的环境区域, 对减少机器人位姿和地图状态变量的不确定性, 避免错误引入地图冗余变量或重复结构至关重要[1−10]. SLAM 中常用激光、雷达、超声波作为外传感器, 随着机器视觉的发展, 利用视觉传感器采集的环境信息和图像处理技术进行场景识别和地图构建, 已成为 SLAM 的重要技术. 基于视觉传感信息的 SLAM 中, 闭环检测主要分为概率计算方法[1−4] 和图像匹配方法[5−10]. 概率计算方法将闭环检测归结为递归贝叶斯估计问题. 首先采用 BoW (Bag-of-words) 等图像建模方法描述机器人每一位置的场景图像, 估计已获取图像与对应位置的先验概率, 对当前时刻, 计算该新场景图像与已访问位置匹配的后验概率, 概

收稿日期 2010-09-21 录用日期 2011-01-22
Manuscript received September 21, 2010; accepted January 22, 2011
国家自然科学基金 (60975015), 中央高校基本科研业务费专项资金 (CDJXS11181162, CDJXS10181133), 教育部博士点基金 (20090191110023), 重庆市重点科技攻关项目 (CSTC2009AB2230) 资助
Supported by National Natural Science Foundation of China (60975015), Fundamental Research Funds for the Central Universities (CDJXS11181162, CDJXS10181133), Doctoral Foundation of Ministry of Education of China (20090191110023), and Key Program for Science and Technology Development of Chongqing (CSTC2009AB2230)
1. 重庆大学计算机学院 重庆 400044 2. 重庆大学软件工程学院 重庆 400044 3. 重庆大学数学与统计学院 重庆 400044
1. College of Computer Science, Chongqing University, Chongqing 400044 2. School of Software Engineering, Chongqing University, Chongqing 400044 3. College of Mathematics and Statistics, Chongqing University, Chongqing 400044

率大于阈值则提取为闭环. 图像匹配方法将视觉闭环检测归结于序列图像匹配问题, 将当前时刻的图像与已获取的图像序列进行相似性匹配, 相似度高于阈值的匹配图像对应了机器人的闭环位置.

如何准确建立场景外观模型, 成为基于外观的视觉闭环检测的关键. 目前, 图像分类中的 BoW 方法被广泛用于视觉 SLAM 及闭环检测[1−8]. Cummins[1] 为克服视觉单词间的独立性假设, 用 Chow-Liu 算法逼近单词的概率分布, 用树形结构的贝叶斯网络生成场景外观模型. Angeli[3] 采用形状和颜色特征, 增量构造视觉单词本, 用贝叶斯方法实时计算闭环概率. 然而, 众多视觉 SLAM 研究往往直接引用了图像分类中的 BoW 方法, 没有考虑 SLAM 中图像处理不同于图像分类的特殊性, 如 SLAM 中图像具有时间连续性特征、不同场景图像间存在视觉混淆现象 (Perceptual aliasing)[1] 等, 尤其随机器人移动, 图像数目急剧增加, 计算负荷随之增大, 然而 SLAM 中需要机器人移动到每一个位置, 都能作出即时响应决策, 这便对图像处理的准确性和实时性提出了更高要求. 虽然有研究[5, 8] 采用了如 k-d tree 等最近邻搜索算法提高从图像特征到视觉单词的投影效率, 但未能从根本上克服单词本的诸多弊端. 同时传统视觉字典本的平面结构制约了单词的表征能力和计算效率, 影响了在 SLAM 中的应用效果. 本文基于树形结构的快速搜索特性, 建立分层的视觉字典树 (Visual vocabulary tree)[6, 11], 不仅视觉单词个数不受限制, 也提高了特征投影时最近邻搜索的效率, 以满足闭环检测的实时性需求.

Callmer 在文献 [6] 中虽然采用视觉字典树描述场景图像, 在城市场景中取得了良好的闭环检测效果, 但如同其他现有研究, 在用得分加权匹配方法计算相似度过程中, 往往通过比较由图像在树中自上而下投影得到的路径节点构成的描述向量, 或只由叶子节点单词构成的描述向量间的距离来计算图像间相似性. 这种方法依然是一种平面结构的匹配模式, 忽略了图像在树中不同层上投影的不同量化差异, 忽略了树的层次节点单词之间的不同表征能力. 本文针对树形结构的分层量化特点提出金字塔匹配方法, 通过深度搜索将图像特征投影到视觉字典树, 计算树节点的 TF-IDF (Term frequency-inverse document frequency) 熵, 利用树的高层节点单词鲁棒性强、闭环检测召回率高, 低层节点单词表征性能好、图像相似性计算准确、闭环检测准确率高的特点, 提出自下而上逐层计算图像间的相似性增量, 最后建立匹配核函数整合相似性增量, 从而避免了传统单一量化尺度的量化误差, 以及基于树的平面匹配方法忽略不同层次视觉单词不同表征性能的不足和边界特征错误分类的累积问题对闭环检测的影响.

由于场景图像的复杂性和图像处理技术本身的计算误差, 在初始匹配图像中存在误匹配, 也即对应了错误闭环, 错误闭环破坏了环境地图构建和机器人定位的准确性. 本文最后采用时间连续性、空间一致性和对极几何约束等后验管理方法[3, 5, 8−9] 剔除错误闭环, 从而提高了闭环检测的准确率.

1 基于视觉字典树的机器人视觉场景描述

移动机器人根据自身运动速度、传感器帧率和实际场景变化等设置视觉传感器的采样频率, 如通过时间域采样[5]、空间域采样[4, 6]、场景内容变化采样[9] 等方法, 采集运动过程的场景图像, 每一图像对应一个机器人位置, 因而 SLAM 中的地图构建和机器人定位以及闭环检测等可归结为一个大规模的具有时间连续性的图像序列的匹配和位置识别问题. 如何准确高效地表征场景图像, 成为视觉 SLAM 系统的基础环节. 直接基于图像低层特征的图像表示方法计算复杂度高, 难以满足 SLAM 的实时性计算需求. 近年, 文本处理的 BoW 方法为机器人位置场景描述提供了一种有效途径. 如图 1, 机器人位置对应拓扑地图中的一个节点, 由传感器采集的该位置处的图像进行描述. 首先提取图像特征, 将特征投影到视觉字典本, 得到图像的视觉单词描述向量, 从而基于向量距离测度计算节点图像间的相似性.

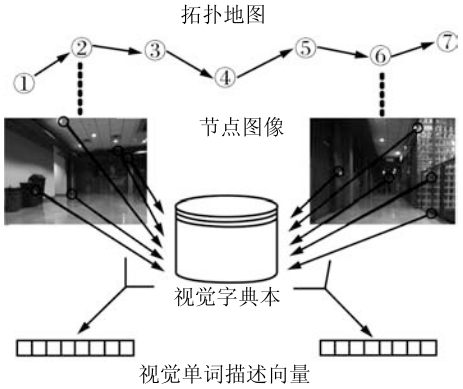


图 1 基于 BoW 的移动机器人位置场景描述

Fig. 1 Scene description of robot′s location based on bag-of-words

1.1 平面视觉字典本

传统视觉字典本的平面结构对单词个数、单词的表征性能、计算代价等都造成极大的制约. 视觉字典本太小, 计算效率高, 但特征数据高度聚类, 生成视觉单词的表征能力下降, 原本不同的特征可能被投影到同一个视觉单词, 导致错误闭环的产生; 视觉字典本增大, 单词的表征能力提高, 但单词鲁棒性

降低, 计算负荷也急剧增加, 导致移动机器人延迟问题. 传统的方法往往依赖经验和实验数据, 折中字典本性能和计算效率, 选取百 ∼ 千量级的视觉字典本, 难以得到最优的应用效果, 尤其在 Web 图像检索、视频处理等实际应用中, 平面结构的视觉字典本难以满足对海量图像的实时性计算需求. 同时平面视觉字典本只在单一尺度下量化连续特征空间, 难以避免量化误差对图像描述的准确性影响.

1.2　基于分层 K-均值聚类的视觉字典树

针对以上问题, 本文采用分层 K-均值聚类建立视觉字典树, 提高视觉单词的表征性能和计算效率. 首先定义树的分支因子 k 和层数 L, 将图像特征划分成 k 个分支, 对每个分支递归地执行 K-均值聚类得到下一层的 k 个更细分支, 直到最大层数 L, 如图 2, 计算每个分支特征的类中心作为树节点, 对应一个视觉单词, 从而生成视觉字典树.

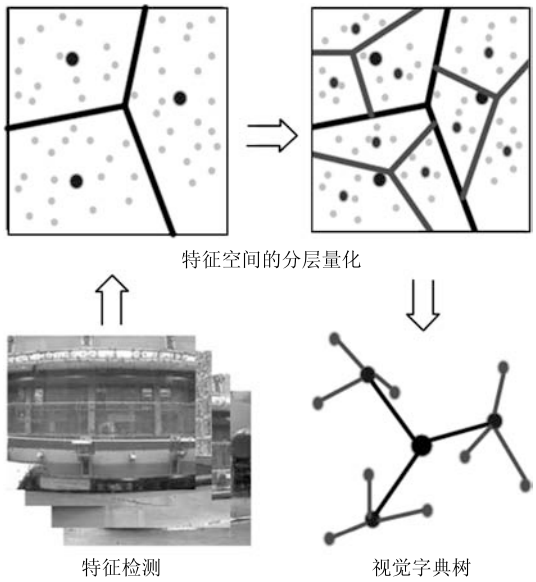


图 2　$k = 3$, $L = 2$ 的视觉字典树的生成流程

Fig. 2　Generation of a simple vocabulary tree with branch factor 3 and only 2 levels

一棵 L 层 k 分支树生成的视觉单词个数为: $\sum_{i=1}^{L} k^i = (k^{L+1} - k)/k - 1 \approx k^L$, 一棵 5 分支树到第 6 层就已具有万级的视觉单词, 其单词空间的表征能力远优于传统视觉字典本的百 ∼ 千级单词空间. 同时, 基于视觉字典树的特征投影, 在每一层只需执行 k 次线性比较, 这样的深度最近邻搜索优于平面视觉字典本的线性搜索, 从而极大提高了算法的效率.

2　分层 TF-IDF 熵的金字塔得分匹配方法

2.1　传统基于视觉字典树的得分相似性计算

基于视觉字典树, 对图像特征进行深度最近邻搜索, 即自上而下逐层与 k 个节点比较, 为特征选择最邻近的视觉单词, 并对树节点计算逆向文档频率作为权值, 采用逆向文档索引存储图像, 再用得分准则 (Score scheme) 进行图像相似度比较[11], 如图 3. 定义每个树节点的权重为 $w_i = \log N/N_i$, N 是待处理图像总数, N_i 是至少有一个特征投影到节点 i 的图像数. 则图像 X 的视觉单词描述向量定义为 $\boldsymbol{X} = (x_i)$, $x_i = n_i w_i$, 并归一化 $\boldsymbol{X} = \boldsymbol{X}/\|\boldsymbol{X}\|$, n_i 表示 $\boldsymbol{X}$ 投影到节点 i 的特征个数. 则在 L_2 范数下两幅图像的相似性得分 S 定义为: $S(X, Y) = \|\boldsymbol{X} - \boldsymbol{Y}\|_2^2 = \sum_i |x_i - y_i|^2 = 2 - 2\sum_{i|x_i \neq 0, y_i \neq 0} x_i y_i$.

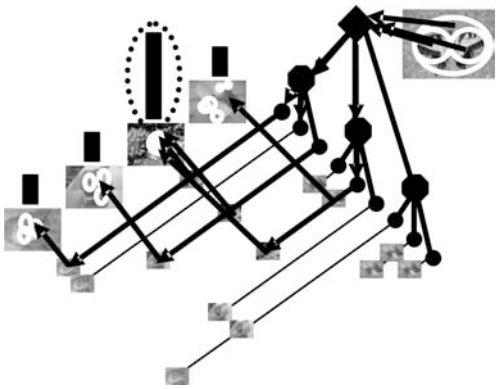


图 3　基于视觉字典树的图像相似性计算

Fig. 3　Image similarity computing based on the visual vocabulary tree

这种方法本质上仍是一种如同视觉字典本中的平面匹配方法, 存在两个固有缺陷: 1) 将所有树节点的视觉单词同等对待, 忽视了不同层次上的单词具有不同的表征能力, 以及图像在不同层上具有不同的相似度的事实. Nister[11] 为减少计算量, 仅用叶子层的节点进行相似性比较. 这些方法没有利用树的层次节点之间相似性的承继关系. 2) 在分层量化过程中, 位于类边界附近的相似特征往往被武断地划分到最近的子树分支中, 这种现象引发的量化误差会逐层累积直到最底层, 从而破坏图像间真实的相似度. 如图 4 中相似特征 p_1 和 p_2 在树的第一层量化空间被投影到视觉单词 C_1, 然而在第二层量化空间却被投影为不同的单词 C_{11} 和 C_{12}, 这种误分一直延续到树的最底层. 利用传统平面模式的匹配方法, 则 p_1 和 p_2 更多地被视为不同特征, 降低了两幅图像的相似性得分, 导致闭环检测的召回率降低.

本文结合分层匹配思想[12], 利用视觉字典树的分层量化特性, 提出了一种基于得分准则的金字塔匹配方法. 通过计算在树的每一层上图像间的相似

性增量, 统计了图像在不同量化尺度的视觉单词空间中的相似程度, 从而有效解决了边界特征等问题, 提高了闭环检测的召回率.

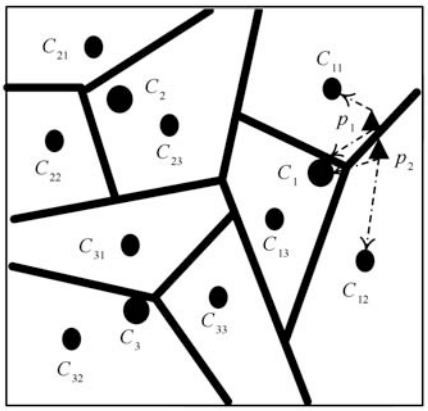


图 4 边界特征的逐层相似度

Fig. 4 The layered similarity of the boundary features

2.2 本文分层 TF-IDF 熵的金字塔得分匹配算法

首先提取图像 X 的 n 个 d 维特征: $X = \{x_1, x_2, \cdots, x_n\}$, $x_i \in \mathbf{R}^d$, 自上而下投影到视觉字典树得到图像 – 单词的逆向索引, 计算图像在每个树节点的 TF-IDF 熵作为图像在该视觉单词的得分权重. 用 $w_i^l(X)$ 表示图像 X 在视觉字典树的第 l, $l \in \{0, 1, \cdots, L\}$ 层的第 i, $i \in \{1, 2, \cdots, k^l\}$ 个节点 O_i^l 处的投影得分, 定义 TF-IDF 熵为

$$w_i^l(X) = \frac{n_i}{n} \log \frac{N}{N_i} \tag{1}$$

图像 X 在整棵视觉字典树中的得分向量记为

$$\boldsymbol{W}(X) = (\boldsymbol{W}^1(X), \boldsymbol{W}^2(X), \cdots, \boldsymbol{W}^L(X)) \tag{2}$$

其中, $\boldsymbol{W}^l(X) = (w_1^l(X), w_2^l(X), \cdots, w_{k^l}^l(X))$ 表示在第 l 层的得分向量. 第 0 层为树根节点, 没有表征能力, 故仅从第一层开始计算.

传统平面模式的匹配方法中, 直接利用 $\boldsymbol{W}(X)$ 与 $\boldsymbol{W}(Y)$ 之间的距离 $\|\boldsymbol{W}(X) - \boldsymbol{W}(Y)\|$ 作为图像 X 与 Y 之间的相似性测量, 或者只计算叶子层的得分向量之间的距离 $\|\boldsymbol{W}^L(X) - \boldsymbol{W}^L(Y)\|$. 这些方法没有利用树结构中得分向量的层次信息, 本文提出金字塔匹配方法在不同层次上计算图像间的相似性, 最后建立核函数整合不同层次的相似性增量.

首先定义图像 X 和 Y 在单个节点 O_i^l 的相似性得分为

$$S_i^l(X, Y) = \min\{w_i^l(X), w_i^l(Y)\} \tag{3}$$

则图像在第 l 层的相似性定义为

$$S^l(X, Y) = \sum_{i=1}^{k^l} S_i^l(X, Y) = \sum_{i=1}^{k^l} \min\left\{w_i^l(X), w_i^l(Y)\right\} \tag{4}$$

视觉字典树由上而下对特征空间由粗到细地划分, 随层数增加, 量化尺度变细, 生成的视觉单词空间对图像刻画越细微, 不同特征投影到同一视觉单词的机率就越小, 故而相似性得分也越小, 这种相似性的逐层继承特性意味着树的上一层空间包含了部分下一层空间的图像相似性, 因此为避免重复累计相似性, 我们从最底层开始由下而上计算图像间相似性增量, 则第 l 层的相似性得分增量 ΔS^l 定义为

$$\Delta S^l(X, Y) = \begin{cases} S^L(X, Y), & l = L \\ S^l(X, Y) - S^{l+1}(X, Y), & 1 \le l < L \end{cases} \tag{5}$$

最后定义金字塔匹配核为

$$K(X, Y) = \sum_{l=1}^{L} \eta_l \Delta S^l(X, Y) \tag{6}$$

参数 η_l 为视觉字典树第 l 层的匹配强度系数, l 越小对特征空间的划分越粗, 不同图像特征投影到同一个视觉单词的约束越小, 因而容易获得较高的匹配得分. 因此本文设计 $\eta_l = 1/k^{L-l}$ 来抑制不同层次的这种匹配差异. 参数 k 大于 1, 值越大, 相似性越依赖于树的最底层的单词空间, 本文实验中取 $k = 2$.

由此, 我们建立了不同图像投影到视觉字典树上的 TF-IDF 熵的金字塔得分匹配方法, 相似性计算的核函数可重写为

$$K(X, Y) = K(\boldsymbol{W}(X), \boldsymbol{W}(Y)) = S^L(X, Y) + \sum_{l=1}^{L-1} \frac{1}{k^{L-l}} (S^l(X, Y) - S^{l+1}(X, Y)) \tag{7}$$

2.3 算法优化计算

由于每幅图像的特征个数远远小于视觉字典树的节点个数, 使得图像投影到字典树上的逆向文档索引具有非常稀疏的结构, 即存在大量节点处的得分 $w_i^l(X) = 0$, 从而在 TD-IDF 得分向量 $\boldsymbol{W}(X)$ 中存在大量的 0 元素. 为提高算法的运行效率, 根据式 (4) 在计算图像的相似性得分时, 不需要计算树的全部节点, 只需计算满足 $w_i^l(X) \neq 0$ 且 $w_i^l(Y) \neq 0$ 的节点 O_i^l 处的相似性得分. 这样的稀疏化处理大大提高了算法的计算效率.

$$K(X,Y)=\sum_{w_i^L(X)\cdot w_i^L(Y)\neq 0}^{i=1:k^L}\min\left\{w_i^L(X),w_i^L(Y)\right\}+\sum_{l=1}^{L-1}\frac{1}{k^{L-l}}\left(\sum_{w_i^l(X)\cdot w_i^l(Y)\neq 0}^{i=1:k^l}\min\left\{w_i^l(X),w_i^l(Y)\right\}-\sum_{w_i^{l+1}(X)\cdot w_i^{l+1}(Y)\neq 0}^{i=1:k^{l+1}}\min\left\{w_i^{l+1}(X),w_i^{l+1}(Y)\right\}\right) \tag{8}$$

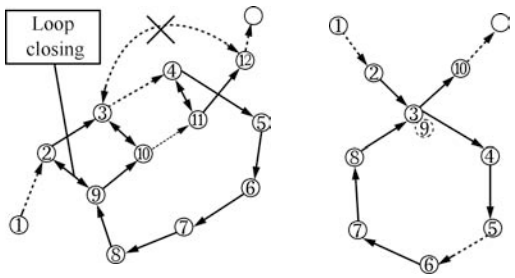
2.4　算法性能分析

本文基于视觉字典树的金字塔得分匹配方法与传统平面匹配方法相比, 优越性如下: 1) 图像视觉单词描述向量的顺序独立性和维数任意性: 传统平面模型的匹配方法大都采用向量距离准则, 只能解决相同维数的有序向量匹配, 使得图像必须投影到相同的视觉单词上构造相同维数的描述向量, 而视觉字典本为提高表征能力往往具有较大的单词个数, 导致图像描述向量维数较高且具有稀疏结构, 而本文得分匹配方法不依赖于距离测度, 对视觉单词是否出现以及出现的顺序没有要求. 2) 匹配的高效性: 本文金字塔匹配算法在捕获图像间不同尺度下的相似性时并没有带来计算复杂度的剧增, 而算法的树形结构大大提高了最近邻搜索的效率. 假设在 d 维空间中, 将一个特征投影到大小为 N 的平面视觉字典本上, 传统线性搜索算法需要与 N 个视觉单词进行相似性比较, 算法复杂度为 $\mathrm{O}(Nd)$, 而本文算法只需在 L 层树的每层中进行 k 次最近邻搜索, 算法时间复杂度为 $\mathrm{O}(kdL)$, 而 $N>>kL$. 3) 多尺度性: 类似于多分辨分析, 在树的低层细粒度划分下, 容易区分图像的细微差异, 然而对背景和噪声敏感; 在树的高层粗粒度划分下, 容易把握图像间的全局相似性, 噪声鲁棒好. 本文金字塔匹配方法综合了不同量化粒度的优点, 在准确计算图像间相似性的同时抑制了背景和噪声的影响. 同时本文方法弥补了单尺度视觉字典本的量化误差和基于视觉字典树的平面匹配方法忽略不同层次单词的不同表征性能的不足.

3　视觉闭环检测及后验处理

当图像间相似性大于给定的全局相似性阈值, 即 $S(X,Y)>T_s$, 提取当前位置与其匹配位置为候选闭环. 候选闭环中往往存在误正闭环, 其产生由诸多原因造成: 1) 视觉单词的歧义性: 将图像压缩到给定大小的视觉单词空间, 不同的特征被投影到相同的视觉单词, 产生一词多义, 加剧了图像的区分难度. 2) 单词间的独立性: 视觉单词方法的词汇独立性假设忽略了图像特征间的空间位置关系和语义相关性, 不同对象的特征可能被识别为相同单词. 3) 视觉混淆现象: 自然场景中存在道路、墙壁、树叶等频繁的特征模式, 这些模式被投影到相同视觉单词上, 加剧了场景间的混淆不清, 如实验结果 (见图 9 和图 11 (a)) 所示.

时间连续性和空间一致性是 SLAM 图像的两个重要特征, 常作为后验处理的约束条件抑制误正闭环. 机器人捕获的场景图像具有时间连续性, 即相邻帧图像对应相同场景的连续变化, 当机器人再次到达已经访问过的位置发生闭环时, 这种闭环位置也应该依次地连续出现一定次数. 如图 5 (a), 当节点 9 与原有节点 2 匹配, 发生闭环, 则后续几帧也将依次发生匹配响应 (节点 10 与 3 匹配, 节点 11 与 4 匹配). 若下一节点 12 与之前 3 匹配, 亦不满足时间连续性特征, 将从闭环候选集中删除, 如实验结果 (见图 11 (a)) 的初始闭环中正好存在这种候选闭环, 采用时间一致性操作可有效排除这种错误闭环的干扰. 机器人为准确捕获场景变化、构建准确地图, 图像采集的帧率一般较小, 故而仅获得一次匹配、发生一次闭环的现象极少, 多为由上述三种原因产生的错误闭环. 如图 5 (b) 仅有一对匹配节点 9 与 3, 这种闭环仅在机器人交叉通过上一次的访问路径, 且场景变化明显、图像采样频率较大的情况下才发生, 此时只需将时间连续性阈值置为 1 即可检测到交叉闭环. 通过设置时间连续性阈值 T_t, 候选闭环中只有连续匹配的帧数满足 $|\{(X_i,Y_j),(X_{i+1},Y_{j+1}),\cdots,(X_{i+1},Y_{j+1})\}|>T_t$ 时, 才被提取为闭环. (X_i,Y_j) 表示图像 X_i 与 Y_j 发生闭环响应, $|\cdot|$ 表示集合元素的个数.



(a) 闭环检测的时间连续性　(b) 一次交叉闭环

(a) The temporal continuity of loop closing　(b) A cross-loop closing

图 5　闭环检测示意图, 节点表示机器人位置, 每个节点由该位置的场景图像表示, 有向边表示机器人移动的顺序

Fig. 5　Illustration of the loop closing by using the topological map: each node described by an environment image indexes a location of the moving robot; edges model the time neighboring relation between the nodes

视觉单词的独立性假设损失了图像原始特征之间的空间关系, 本文引入一种低层特征空间位置的

一致性约束抑制误正闭环. 闭环的发生意味着机器人再次访问到相同场景, 则同一场景特征的某些几何结构关系会保持不变, 最近邻关系可视为一种简单的空间一致性约束. 将图像特征及其空间位置上的最近邻点投影到视觉单词空间, 则图像可由两个向量描述, 一是每个特征对应的视觉单词列表 v, 另一个是每个特征的空间最近邻点对应的单词列表 v_{nn}. 因为机器人在不同时刻采集的同一场景图像, 所有匹配特征在两幅图像中的空间最近邻关系都不相同的情况极少, 若保持这种最近邻关系的匹配特征点对少于阈值, 即使两幅图像具有较高的相似性得分, 但它们来自同一场景的可能性也极小. 定义 $\varepsilon = |v^1_{nn} \cap v^2_{nn}|/|v^1 \cap v^2|$, 即两幅图像具有相同最近邻的特征个数与匹配特征个数的比值, 空间一致性约束定义为候选闭环的 ε 小于阈值 T_ε 则予以剔除.

SLAM 中视觉闭环图像是对同一场景的不同视角成像, 多视角几何算法常被用于确认初始匹配[3]. 闭环处的两幅图像应满足对极几何约束[13], 用 RANSAC 算法根据闭环图像的特征匹配点对估计两幅图像间的基础矩阵, 若满足基础矩阵的内点数比例 $\rho = inliers/\min(n_X, n_Y)$ 超过给定阈值 T_ρ, 则保留该两幅图像位置为正确闭环, n_X, n_Y 分别表示图像 X, Y 的特征个数.

三种后验确认操作从不同角度尽可能排除了错误闭环的干扰. 如实验结果 (见图 11 (a)) 中, 存在相邻两帧都与已访问的同一帧匹配的闭环情况, 由于相邻两帧皆是对同一场景的成像, 只是位置上有微小差异, 则用空间一致性和对极几何约束都难以排除这种情况, 而时间一致性约束正好可以剔除其中一个闭环响应. 阈值 T_s, T_t, T_ε, T_ρ 的选取往往折中于闭环检测的准确率和召回率, 保证较高的准确率往往会牺牲召回率, 反之亦然, 实践中需要根据不同的场景和准确性需求设定适合的值. 本文首先通过设置较小的全局相似性阈值 T_s (本实验中 $T_s = 0.4$), 使初始闭环中尽可能包含所有真实闭环, 以获得较高的召回率, 再设置合适的后验阈值剔除错误闭环, 提高闭环准确率 (本实验中获得最佳检测结果时的阈值选取为 $T_t = 8$, $T_\varepsilon = 0.2$, $T_\rho = 0.1$).

4 实验结果及分析

本实验用 iRoomba 作为移动机器人, 视觉传感器采用 Monocular Wi-Fi camera, 在 Matlab 平台上集成 iRoomba 工具包和无线摄像头的控制程序, 根据 iRoomba 返回的 odometry 值 (三维数据: 机器人位置的二维坐标和朝向) 绘制机器人真实运动轨迹, 根据摄像头采集的机器人每一时刻的场景图像进行闭环检测. iRoomba 在某一实验中心约 35 米 × 10 米的环形走廊运行一周再回到出发点, 继续运行产生约 20 米的闭环轨迹, 如图 6, 其中图像 1 至 6 是不同位置的场景示例, 匹配图像 1 与 5, 2 与 6 代表了机器人的闭环响应.

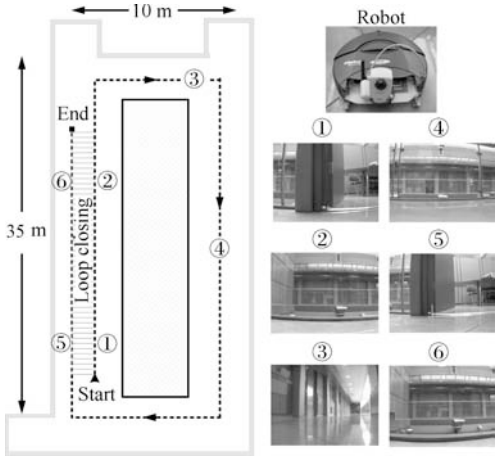


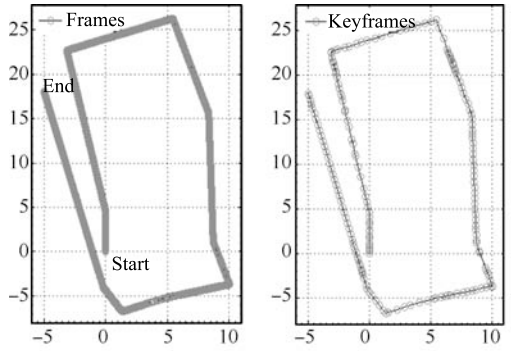
图 6 机器人闭环运行轨迹及典型的走廊场景图像, 闭环示例如 1 与 5, 2 与 6

Fig. 6 Overall robot trajectory for the indoor environment and the corridor image examples. Loop closures are designed to arise on the left corridor, for example, locations 1 and 5, 2 and 6 are loop closing

设置摄像头采样频率约为 0.2321 s, 共采集 3 877 张 640 像素 × 480 像素的彩色图像. 图 7 (a) 是机器人运行的实际轨迹以及在每个位置的场景图像. 高采样频率可以获取场景的准确信息, 但带来巨大的计算量. 关键帧提取是视频处理中有效的数据压缩方法, 本实验用基于 SIFT 特征匹配的图像相似性计算方法, 检测连续图像内容变化大于阈值的图像作为关键帧, 获取 193 张关键帧图像作为机器人运行环境的位置标识. 如图 7 (b), 关键帧趋于均匀地分布在整个运行轨道上.

4.1 基于视觉字典树的计算效率

基于视觉单词的闭环检测中, 为每个图像特征搜索最相似的视觉单词是影响实时闭环检测效率的关键, 本实验对比了树形结构的最近邻搜索与传统平面结构的线性搜索的效率. 图 7 是在 193 张场景图像上的 149 112 个 128 维 SIFT 特征的平均最近邻搜索时间对比实验. 基于视觉字典树的计算效率由分支数和层数决定, 图 8 (a) 中设置分支数为 3, 随树的层数从 4 增大到 9, 对应叶子节点代表的视觉单词数从 3^4 增大到 3^9, 远远大于实验中视觉字典本的单词数: 50, 100, 150, 200, 250 和 300. 实验 7 (b) 设置层数为 5, 分支数从 3 增大到 8, 对应的视觉单词数从 3^5 增大到 8^5, 视觉字典本的单词数取值

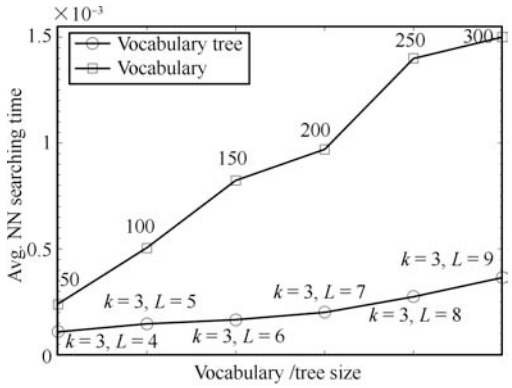


(a) 原始帧分布　(b) 本文提取的关键帧分布

(a) The distribution of original frames　(b) The distribution of keyframes extracted from the frames on the trajectory

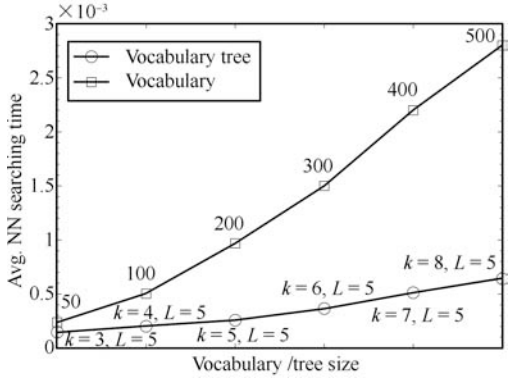
图 7　机器人运行轨迹及摄像头采集的场景图像分布

Fig. 7　The trajectory of iRoomba robot and frames captured by the camera



(a) 分支数 $k = 3$, 比较不同层数的视觉字典树与不同大小的视觉字典本的平均最近邻搜索时间

(a) The vocabulary tree with fixed branch factor 3 and varying levels



(b) 层数 $L = 5$, 比较不同分支数的视觉字典树与不同大小视觉字典本的最近邻搜索效率

(b) The vocabulary tree with fixed levels 5 and varying branch factor

图 8　最近邻搜索效率对比

Fig. 8　Comparison of average nearest neighbor search time between the vocabularies with increasing size

为 50, 100, 200, 300, 400 和 500. 从图 8 (a) 和图 8 (b) 的对比结果可以看出, 视觉字典树的单词个数随分支数和层数的增加而剧增, 且最近搜索速度远远优于传统平面单词本的线性搜索. 故而本文算法更能满足机器人大环境闭环检测的实时性需求.

4.2　平面匹配的量化误差

为说明单一尺度下视觉单词空间对视觉特征的量化误差, 首先用线性搜索在 149 112 个特征集中计算每个特征的最近邻点, 其中互为最近邻的特征点对个数记为 P_{NN}, 计算某一尺度下, P_{NN} 个最近邻点对被投影到不同视觉单词的个数为 P^*_{NN}, 定义量化误差 $\Delta = P^*_{NN}/P_{NN}$. 表 1 统计了不同尺度下的量化误差, 取视觉字典树的分支数为 5, 随层数从 3 增加到 8, 量化粒度越细, 生成视觉单词的表征性能越好, 但被错分的边界点增多, 量化误差也累积增大. 所以无论是传统基于单一量化尺度的视觉字典本还是基于视觉字典树的平面匹配方法都无法克服量化误差对最近邻搜索和相似性计算的影响, 本文提出的金字塔匹配方法自下而上统计不同尺度下的相似性增量, 有效减小了累积量化误差对相似性计算的干扰.

表 1　不同量化尺度下的量化误差

Table 1　The quantization errors at different scales of vocabulary tree

字典树的层数	3	4	5	6	7	8
叶子层单词数	5^3	5^4	5^5	5^6	5^7	5^8
量化误差 (%)	17.45	23.37	36.78	45.28	49.53	59.43

图像间相似性计算是闭环检测的关键步骤, 本实验通过比较不同算法的相似性矩阵与基准相似性矩阵的差异, 定量评估平面匹配与本文分层匹配的性能. 逐点一一匹配的方法复杂度高, 但其最近邻搜索的准确性是其他近似最近邻搜索算法无可比拟的, 故用特征点匹配方法计算 193 张图像之间的相似性得到基准相似性矩阵, 如图 9. 在同一视觉字典树中, 分别计算传统平面匹配方法和本文分层匹配方法的相似性矩阵, 同时生成与树的叶子节点数相同的视觉字典本, 计算相似性矩阵. 图 10 是三种方法生成的相似性矩阵与基准相似性矩阵的差异比较. 视觉字典树的分支因子取值为 5, 层数取值 3、4、5, 对应叶子单词个数为 125、625、3 125, 将平面视觉字典本大小依次定为相同值. 虽然随单词个数的增长, 基于视觉字典本方法的效果随之改善, 但如图 8 所示, 算法的计算时间也线性增加, 而本文分层匹配方法不仅计算效率高, 而且获得与基准相似性矩阵最为接近的结果.

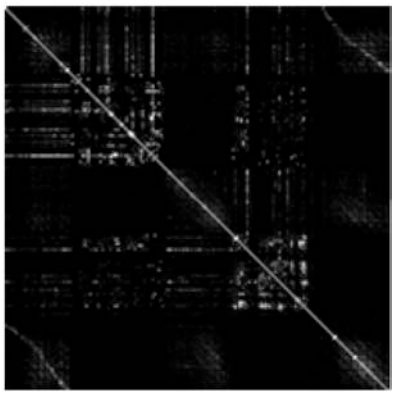


图 9 基于特征匹配的视觉相似性矩阵

Fig. 9 Visual similarity matrix by using feature matching

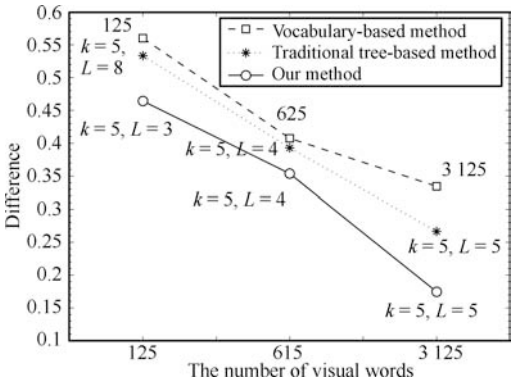


图 10 三种算法计算相似性矩阵与基准矩阵的误差比较

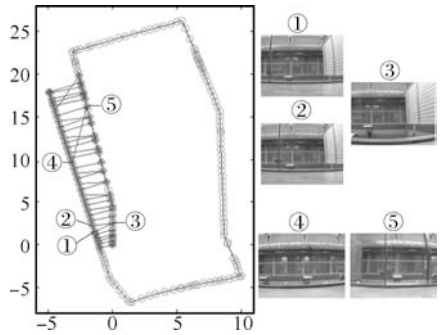
Fig. 10 Comparison of the differences between similarity matrixes and groundtruth

4.3 闭环检测的错误率和召回率比较

如图 9, 相似性矩阵中高亮的非主对角线正对应了机器人运行的闭环, 实验中观察闭环仅发生在最后一段轨迹中, 所以只从第 150 个关键帧图像开始计算闭环的错误率和召回率作为评估指标, 定量评估不同闭环检测方法的性能.

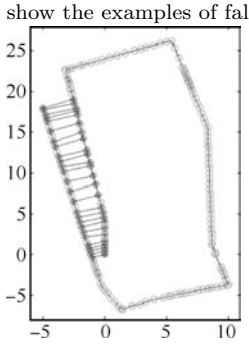
首先在图 9 的基准相似性矩阵中, 基于全局相似性阈值得到如图 11 (a) 的候选闭环, 可见候选闭环中存在错误闭环. 如位置 1 和 2 处的同一场景图像仅有微小差异, 同时与位置 3 发生闭环响应, 皆满足空间一致性和对极几何约束, 故本文仅保留其中相似性较大者为真实闭环. 位置 4 与 5 是机器人在不同位置对不同场景的成像, 由于视觉混淆引起该两幅图像发生较强的匹配响应, 采用时间一致性可有效排除此类错误闭环. 图 11 (b) 是利用后验管理得到的真实闭环, 将该 26 个闭环响应作为本实验场景的基准闭环.

在视觉单词个数、相似性阈值、后验管理参数等的不同取值下, 运行不同算法检测闭环, 基于基准闭环统计每次检测结果的准确率和召回率. 图 12 是基于本文视觉字典树的金字塔匹配方法、基于视觉字典树投影路径的平面匹配方法和基于平面视觉字典本方法所获得的 P-R 曲线, 本文算法在保证 100 % 准确率的条件下可得到约 56 % 的召回率, 而基于视觉字典树投影路径的平面匹配方法约为 53 %, 基于平面视觉字典本方法仅为 41 %, 可见本文算法获得了最佳的闭环检测结果.



(a) 全局相似性阈值提取的候选闭环, 1 ∼ 5 代表错误闭环示例

(a) Loop closure candidates based on global similarity threshold, 1 ∼ 5 show the examples of false loop closures



(b) 经后验管理提取的真实闭环

(b) The true loop closures selected by using posteriori management

图 11 基于特征匹配的闭环检测

Fig. 11 Loop closure detection by feature matching

5 结论

场景图像描述及相似性计算是视觉 SLAM 及机器人闭环检测的关键问题, 本文提出了一种基于视觉字典树的金字塔图像匹配的视觉闭环检测方法. 视觉字典树克服了传统平面视觉字典本性能受制于有限单词个数的不足, 提高了视觉单词空间的表征能力; 其树形结构为图像特征投影提供了高效的最近邻搜索结构, 满足闭环检测的实时性需求; 金字塔匹配方法通过建立核函数整合不同量化尺度下图像间的相似性增量, 有效消除了单一尺度下的量化误差对闭环检测的影响, 以及弥补了传统视觉字典树中基于投影路径的平面匹配模式中, 忽略不同层视觉单词具有不同区分度的缺陷; 后验确认操作有效抑制了误正闭环的干扰. 在实际移动机器人闭环检

测实验中, 本文算法提高了视觉闭环的检测效率和准确率.

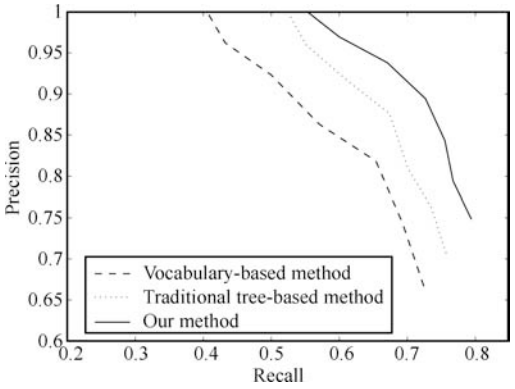


图 12 三种不同闭环检测算法的准确率和召回率曲线

Fig. 12 Precision-recall curves of loop closure detection of the three variant methods

References

1 Cummins M, Newman P. Probabilistic appearance based navigation and loop closing. In: Proceedings of the IEEE International Conference on Robotics and Automation. Rome, Italy: IEEE, 2007. 2042−2048

2 Bazeille S, Filliat D. Combining odometry and visual loop-closure detection for consistent topo-metrical mapping. *RAIRO Operations Research*, 2010, **44**(4): 365−377

3 Angeli A, Filliat D, Doncieux S, Meyer J A. Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, 2008, **24**(5): 1027−1037

4 Cummins M, Newman P. FAB-MAP: probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 2008, **27**(6): 647−665

5 Ho K L, Newman P. Loop closure detection in SLAM by combining visual and spatial appearance. *Robotics and Autonomous Systems*, 2006, **54**(9): 740−749

6 Callmer J, Granstrm K, Nieto J, Ramos F. Tree of words for visual loop closure detection in urban SLAM. In: Proceedings of the Australasian Conference on Robotics and Automation. Canberra, Australia. 2008. 1−8

7 Williams B, Cummins M, Neira J, Newman P, Reid I, Tardos J. An image-to-map loop closing method for monocular SLAM. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Nice, France: IEEE, 2008. 2053−2059

8 Ho K L, Newman P. Detecting loop closure with scene sequences. *International Journal of Computer Vision*, 2007, **74**(3): 261−286

9 Kim J, Kweon I S. Robust feature matching for loop closing and localization. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. San Diego, USA: IEEE, 2007. 3905−3910

10 Zhao Feng-Da, Kong Ling-Fu. An approach to loop-closing based on images matching. *Journal of Yanshan University*, 2008, **32**(2): 115−119
(赵逢达, 孔令富. 一种基于图像匹配的闭环检测方法. 燕山大学学报, 2008, **32**(2): 115−119)

11 Nister D, Stewenius H. Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE, 2006. 2161−2168

12 Grauman K, Darrell T. The pyramid match kernel: discriminative classification with sets of image features. In: Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005. 1458−1465

13 Nister D. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(6): 756−770

李 博 重庆大学计算机学院博士研究生. 2007 年获重庆大学数理与统计学院硕士学位. 主要研究方向为图像处理、模式识别和机器人视觉闭环检测. 本文通信作者. E-mail: boli.cqu@gmail.com
(**LI Bo** Ph. D. candidate at the College of Computer Science, Chongqing University. He received his master degree from Chongqing University in 2007. His research interest covers image processing, pattern recognition, and visual loop closure detection. Corresponding author of this paper.)

杨 丹 重庆大学计算机学院教授. 主要研究方向为图像处理、机器视觉、人工智能和软件工程.
E-mail: dyang@cqu.edu.cn
(**YANG Dan** Professor at the School of Software Engineering, Chongqing University. His research interest covers image processing, machine vision, artificial intelligence, and software engineering.)

邓 林 重庆大学数理与统计学院讲师. 2005 年获重庆大学硕士学位. 主要研究方向为图像处理和模式识别.
E-mail: lin.denglinlan@gmail.com
(**DENG Lin** Lecturer at the College of Mathematics and Statistics, Chongqing University. She received her master degree from Chongqing University in 2005. Her research interest covers image processing and pattern recognition.)