doi:10.19306/j.cnki.2095-8110.2019.04.010

单目相机物体位姿估计方法研究

邢加伟,田海峰,王　芳

(航天科工智能机器人有限责任公司,北京 100074)

摘　要:现有的机器视觉通常以边缘轮廓和角点作为特征,因此要求背景单一,对环境结构化依赖程度高。为了拓展机器人的应用范围,使其脱离结构化的环境,提出了一种基于 SIFT 特征点和 PNP 技术的单目相机估计目标物体位姿的方法。以 BumbleBee 双目相机为硬件基础,以 C++为开发平台,结合了 Eigen 计算库、OpenCV 图像处理库和 Triclops 库,开发了单目视觉位姿估计算法,实现在复杂背景下对表面纹理较为丰富的物体的位姿估计。利用试验对所提方法进行了验证,试验结果表明,该算法具有较高的估计精度,可以作为机器抓取的依据。

关键词:机器视觉;视觉伺服系统;位姿估计;尺度不变特征转换;透视 n 点定位;随机抽样一致性

中图分类号:TP242.62　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:

文章编号:2095-8110(2019)04-0071-07

Study on Object Position and Pose Estimation Method of Monocular Camera

XING Jia-wei, TIAN Hai-feng, WANG Fang

(Aerospace Science & Industry Intelligent Robot Co., Ltd., Beijing 100074, China)

Abstract: The traditional methods of machine vision are usually characterized by edge contour and corner points, so it requires a monotonous background and relies on the structured environment strongly. In order to expand the application scope of robot and be away from structured environment, a method based on SIFT (Scale-Invariant Feature Transform) feature points and PNP (Perspective-n-Point) technique for monocular camera estimation of object position and pose is proposed. Taking the BumbleBee binocular camera as the hardware basis, C++ as the development platform, combined with Eigen computing library, OpenCV image processing library and Triclops library, the monocular vision pose estimation algorithm is developed to realize the pose estimation of objects with rich texture under complex background. The proposed method is verified by experiments, and the experiments results shows that the method can estimate with high accuracy and can be used as a basis for robot grasp.

Key words: Machine vision; Visual servo system; Position and pose estimate; SIFT; PNP; Random sampling consistency

0　引言

计算机视觉技术在目标识别和目标位姿求取方面具有强大的优势。人类获取的信息有 70%都是由视觉获取。视觉具有信息量大、精度高、非接触和响应快等特点。

收稿日期:2018-10-16;**修订日期**:2018-12-12

作者简介:邢加伟(1994-),男,硕士研究生,研究方向为计算机视觉。E-mail:staxjw@163.com

目前，用于工业流水线的视觉伺服基本都是基于轮廓和角点等特征在背景单一的结构化环境下完成检测等功能。一旦脱离结构化环境，在复杂环境背景下，轮廓和角点等特征鲁棒性会大幅下降，基本失去作用。例如，工业装配流水线的视觉抓取系统[1]，以单目相机作为传感器，利用边缘轮廓的几何外形作为特征，需要被抓物体离散放置，相邻物体有间隙，每个物体放置朝向需要一致。这种解决方案在流水线上效率非常高，但是，其限制条件决定了其只能在工业流水线等结构化环境中使用。基于轮廓的定位由于其本身的特点，对于位姿的确定存在一定的缺失，通常不能完全确定 6 个自由度的信息。

对于应用场景为复杂环境下的机器人，例如特种机器人、排爆机器人和服务机器人等，技术仍未成熟。不同于工业机器人在结构化环境下对工件的抓取，排爆机器人等在复杂环境下的智能抓取面临着诸多挑战，例如动态化环境、光照变化、几十乃至上百种目标物体、复杂背景、物体间的相互遮挡等。对于机器人按照用户命令抓取指定物体的应用情境，一个完整的抓取过程一般包括物体识别、位姿计算、抓取姿态生成、运动规划与执行等多个环节。其中，对目标物体的识别与定位是抓取成功的前提条件。本文针对抓取中的核心问题——位姿确定，提出了一种基于特征点匹配和透视 n 点定位(Perspective-n-point，PNP)技术的方法，可以使机器人适合非结构化环境，拓展其任务范围。特征点匹配在视觉导航领域已经广泛应用[2-5]，证明了其可靠性与实用性。对于表面存在纹理的物体，不论形状规则与否，在利用深度相机或者双目相机进行建模后，对消费级相机进行张正友标定[6]，即可进行位姿估计，鲁棒性和实时性均满足物体抓取的需求。

1 算法原理

1.1 算法总述

本文提出的位姿识别算法，需要在物体上固连一个假想的坐标系(下文简称物体坐标系)，通过围绕物体扫描一周，将扫描得到的物体表面进行尺度不变特征变换(Scale-Invariant Feature Transform，SIFT)特征点提取，利用双目相机计算出特征点在相机坐标系下的坐标，并根据相机在扫描物体时相对于物体的位姿，将特征点在相机坐标系下的坐标转换为其在物体坐标系下的坐标，得到不同视角下物体可视面的 SIFT 特征点稀疏点云(后文简称模型点云)。在算法进行位姿估计时，拍摄一张物体照片，找到与其最接近的视角下的模型点云。通过 SIFT 特征匹配，得到相对应的物体坐标系 3D 坐标队列和图像坐标系 2D 坐标队列。利用 PNP 算法和 RANSAC 算法，求解出物体坐标系相对于相机坐标系的位姿。

1.2 SIFT 特征点

SIFT[7]全称为尺度不变特征转换 。它用来侦测与描述影像中的局部性特征，具有旋转不变性(拍摄角度)、尺度不变性(拍摄距离)等优势，而且对于视角倾斜、光照、噪音影响，目标遮挡和光照变化具有一定的鲁棒性。通过旋转照相机拍摄同一物体不同角度的多张图像时，这些图像的对应点的 SIFT 描述符具有很高的相似度。基于这些特性，因为它们是高度显著而且相对容易撷取，在母数庞大的特征数据库中，很容易辨识物体而且鲜有误认。经过优化和 GPU 加速的 SIFT 特征点提取匹配可以达到实时的效果。而且扩展性好，可以方便地与其他的特征进行联合，例如 ORB 特征[8]、SURF 特征[9]、Harris 角点[10]等。

SIFT 使用流程分解为如下五步：

1)尺度空间极值检测

首先构建差分高斯金字塔(图 1)，高斯差分金字塔由高斯金字塔差分获得，高斯金字塔模型是将原始图像不断降阶采样和高斯滤波，得到一系列大小不一的图像，由大到小、从下到上构成的塔状模型。原图像为金子塔的第一层，每次降采样所得到的新图像为金字塔的一层，每层的一张图像使用不同参数做高斯模糊，使得金字塔的每层含有多张高斯模糊图像。将金字塔每层多张图像合称为一组(Octave)，金字塔每层只有一组图像，组数和金字塔层数相等。金字塔的层数根据图像的原始大小和塔顶图像的大小共同决定。最后，通过搜索差分金字塔，将每一个点与其上下左右共 26 个点作比较，找到极值点。

2)关键点定位

因为其在金字塔中的坐标为离散值，可以通过泰勒二阶展开三维函数得到精细化的(x,y,δ)即为 SIFT 特征点的坐标，其中 x、y 为在图像中的位置，δ 为当前所在的尺度值。因为将极值点所在的尺度值 δ 作为变量求出，所以具有尺度不变性。

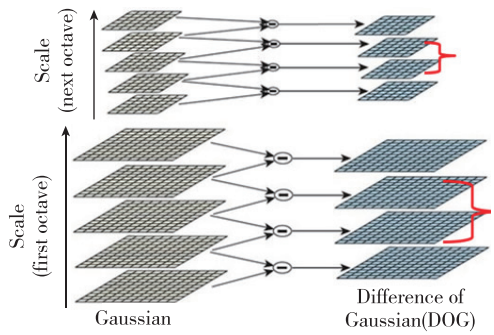

图 1　高斯差分金字塔示意图

Fig. 1　Gaussian difference pyramid diagram

3)方向确定

为了使描述符具有旋转不变性，需要利用图像的局部特征给每一个关键点分配一个基准方向。采集其所在高斯金字塔图像 3δ 邻域窗口内像素的梯度和方向分布特征。

$$m(x,y)=\sqrt{(L(x,y+1)-L(x,y-1))^2+(L(x+1,y)-L(x-1,y))^2} \tag{1}$$

$$\theta(x,y)=\tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y))) \tag{2}$$

其中，L 为关键点所在尺度空间的灰度分布，$m(x,y)$ 为梯度模值，$\theta(x,y)$ 为梯度的方向。使用直方图统计邻域内像素的梯度和方向。梯度直方图将 0°～360°的方向范围分为 36 个柱(bins)，其中每个柱 10°，每个柱高度为这个方向的模值之和。直方图的峰值方向代表了关键点的主方向。

4)关键点描述

以主方向为 x 轴，建立坐标系，在其尺度空间上将其邻域划分为 4×4 个小区域。每个小区域内若干像素，统计每个像素的梯度，将 360°分为 8 个方向，统计每个方向梯度值之和做分布直方图，可得到一个 4×4×8=128 维的向量，就是此处 SIFT 点的描述子。

5)SIFT 特征点匹配

2 个 SIFT 特征点描述队列，对于第一个队列中的某一个特征点，在第二个队列中找到与其欧氏距离最近的 2 个 SIFT 特征点，如果最近的点距离/次近的点的距离小于某个比例值(通常取 0.4～0.6)，则认为最近的那个点是匹配点。

1.3　PNP 算法

PNP 是求解 3D 到 2D 点对运动的方法，描述了当知道 n 个 3D 空间点及其 2D 投影位置时，如何估计相机和空间的相对位姿。

PNP 问题有很多种求解方法，例如，只有 3 对点估计位姿的 P3P[11]、直接线性变换(Direct Linear Transformation，DLT)、高效 PNP[12](Efficient PNP，EPNP)、直接最小二乘法[13](Direct Least-Squares，DLS)等。

其中，DLT 最少需要 6 对特征点可实现变换矩阵的求解。但是将变换矩阵看成了 12 个未知数，忽略了他们之间的联系，因为旋转矩阵 $\boldsymbol{R}\in SO(3)$，用直接线性变换求出来的解未必满足该约束条件。

P3P 只需要 3 对匹配点即可求解出相对位姿，但是匹配点通常多于 3 对，P3P 并不能有效利用，而且 P3P 算法当受到噪声影响时算法失效，鲁棒性较差。

EPNP 算法的主要思想是选取世界坐标下的 4 个控制点坐标为 $\boldsymbol{C}_w=[0,0,0,1]^{\mathrm{T}}$，$[1,0,0,1]^{\mathrm{T}}$，$[0,1,0,1]^{\mathrm{T}}$，$[0,0,1,1]^{\mathrm{T}}$；通过 n 个 3D 点在相机平面的投影关系，以及与这 4 个控制点的权重关系，构建一个 12×12 方阵，求得其零空间特征向量，可以得到虚拟控制点的相机平面坐标，然后使用 POSIT 算法即可求出相机位姿。

DLS 算法的主要思想是对方向向量而不是位置向量进行最小二乘求解，同时加入了噪声的考虑。对比 EPNP 算法，DLS 算法在匹配点对较少的情况下表现更优。

综上所述，因为匹配点通常远远大于 4 对，但偶尔存在匹配较少的情况，所以本文选择 DLS 算法。

1.4　随机抽样一致性算法

随机抽样检验一致性[14](Random Sample Consensus，RANSAC)算法，可以从一组包含局外点的观测数据集中，通过迭代方式估计数学模型的参数。它是一种不确定的算法——它有一定的概率得出一个合理的结果；为了提高概率必须提高迭代次数。

RANSAC 的基本假设是：

1)数据由局内点组成，例如：数据的分布可以用一些模型参数来解释；

2)局外点是不能适应该模型的数据；

3)除此之外的数据属于噪声。

其中局外点产生的原因有：噪声的极值、错误的测量方法、对数据的错误假设等。

PNP 算法与 RAMSAC 算法结合的工作流程为：

1)随机从样本中选择 5 个点对,利用 PNP 算法算出外参数矩阵。

2)用 1)中的外参数矩阵去测试样本中的其他点对,即用 3D 点经过外参矩阵计算投影出的 2D 坐标和实际的 2D 坐标比较,根据其是否小于阈值,确定点对为局内点还是局外点。

3)如果局内点足够多,并且局内点多于原有最佳外参矩阵的局内点,那么将这次迭代的外参矩阵设为最佳外参矩阵。

4)重复 1)~3)步直到找到最佳外参矩阵。

2 位姿算法实现

2.1 建立模型

建立模型时,为了使物体在双目相机视场的重叠面积尽可能大而且便于计算,首先将物体坐标系和双目左相机坐标系重合,将物体坐标系原点沿物体坐标系 X 轴正方向平移到双目相机基线(双目相机 2 个光心距连线)中点,平移距离为 $b/2$;再沿物体坐标系 Z 轴正方向移动到照片相机可以拍摄到物体全貌,距离记为 d。为了体现算法的通用性,模型选择为普通食品盒。通过将物体沿物体坐标系 Y 轴旋转拍摄来模拟相机绕物体旋转拍摄,并利用双目相机拍摄物体,得到左右 2 张图片。若每次旋转弧度过大,则 SIFT 特征点匹配成功率会严重下降;若每次旋转弧度过小,则需要拍摄的图片数量过多,增加计算量,影响实时性。因此,经过实验,选择每次正向转动 $\pi/20$ 弧度,共拍摄 40 张照片。对每张照片加入包围盒,以仅利用照片中物体表面图像的 SIFT 特征点,将其他的特征点舍弃。将左右 2 张图片进行 SIFT 特征点提取并计算其描述子。SIFT 提取时,选择层数根据如下公式

$$O=\log_2\min(M,N)-2 \tag{3}$$

其中,M、N 分别为图片在横向和纵向 2 个方向的像素数。

摄像机输出的图像分辨率为 1024×768,即 $M=1024$,$N=768$,计算得到层数应为 7.5,四舍五入选择 8 层。高斯模糊系数初始值取 1.6,其余参数采用 OpenCV 默认值。匹配 2 幅图中 SIFT 特征点,将匹配的点对利用双目视差法[15]算出特征点在双目相机左相机坐标系下的坐标。根据双目相机左相机坐标系和物体坐标系的位姿关系,将 SIFT 特征点坐标从双目相机左相机坐标系转换到物体坐标系下,转换公式如下

$$\boldsymbol{P}_{\mathrm{obj}}=\boldsymbol{R}_i^{*}(\boldsymbol{P}_{\mathrm{c}}+[0,b/2,d]^{\mathrm{T}}) \tag{4}$$

其中,$\boldsymbol{P}_{\mathrm{obj}}$ 表示点在物体坐标系下的坐标,$\boldsymbol{P}_{\mathrm{c}}$ 表示对应的点在双目相机左相机坐标系下的坐标,$[0,b/2,d]^{\mathrm{T}}$ 是物体坐标系原点相对于双目相机左相机坐标系的位置,$\boldsymbol{R}_i$ 是拍摄的第 i 帧相对于第 1 帧的姿态旋转矩阵,在每次旋转弧度为 $\frac{\pi}{20}$ 时,其对应的旋转向量为 $[0,-\frac{\pi}{20}\times i,0]$。

将每一帧的左相机拍摄的图片进行 SIFT 特征点提取,因为每一帧图片能够提取的 SIFT 特征点均为多个,所以把这些特征点存入堆栈之中,提取的描述子和坐标转换后的坐标同样存入堆栈之中,最后存入模型库。对应转动 40 次,模型库共包括 40 包数据,分别为 data1,data2,…,data40。每包数据主要包括 SIFT 特征点堆栈(vector<keypoints>)、SIFT 描述子堆栈(vector<descriptor>)和特征点在物体坐标系中的坐标堆栈(vector<Point3f>)这 3 个数组。至此,模型建立完毕。

2.2 与模型匹配

拍摄一张包含物体的照片(下文中简称实验照片)并测量此时相机相对于物体的位姿,对实验照片提取 SIFT 特征点和计算描述子,并与模型库中每包数据进行特征点匹配。因为每个特征点的描述子为一个 128 维向量,所以每一个匹配点对的暴力匹配的结果为一个描述子的点乘,即一个欧氏距离。每一个匹配结果为一个欧式距离(distance)堆栈(vector<>),共得到 40 个匹配结果即 vector1<distance>, vector2<distance>, …, vector40<distance>。根据实验,即使待测照片和模型照片重合度非常高,也会出现误匹配的情况,因为匹配对欧氏距离越大,误匹配几率越大。为了增强算法实时性,消除误匹配的干扰,因此,将每个匹配结果的欧式距离堆栈从小到大排列,得到 40 个经过排序的欧式距离序列,取其前 80%以尽可能消除误匹配的影响,并计算剩余欧氏距离的平均值,平均值最小的模型包就是匹配最好的模型包。

2.3 求取位姿

为了提高实时性,PNP 解算不利用全部的匹配点,而是利用 3.2 节中的排序结果,求取匹配最好的模型包对应的欧氏距离序列的最小值 $\mathrm{distance}_{\min}$,将大于 $n*\mathrm{distance}_{\min}$ 的进行舍弃,选择倍数 n 通过实验确定。利用 PNP 算法和 RANSAC 算法进行

误匹配滤除和位姿解算，RANSAC 算法迭代次数选择为 20000(只为了得到最好结果，实际工作远远小于这个迭代次数就会得到最佳效果)，通过实验确定数据是否适用于模型的阈值。算出模型本体坐标系相对于相机坐标系的旋转向量和平移向量。

3　实验

3.1　算法实现平台

本文所采用的视觉硬件为加拿大 PointGray 公司的 BumbleBee BB2 双目相机，双目仅在建立模型时使用，在位姿确定时仅用左镜头图像。实验平台(图 2)搭建在普通实验室中，光照条件变化较大，可以测试系统对光照的鲁棒性。计算设备采用第四代英特尔 i7 处理器，16G 内存。模型搭载平台为三轴旋转平台。

图 2　试验验证平台

Fig. 2　Experiment platform

软件平台为 Ubuntu16.04 下 QT 开发环境。利用了 Eigen 数学计算库、OpenCV 图形处理库、BumbleBee 自带的 triclops 双目视觉库。

受实验条件限制，初始化对齐过程不可避免存在误差，为了最大限度消除误差，测量其 3 个自由度上的旋转幅值来计算其精确度。每次绕一个轴旋转 10°，通过单目视觉位姿算法测量其转动角度，相减得到其误差值。

3.2　实验设计

通过实验确定如下参数：

1)输入 PNP 的特征点选择倍数 n，通常在 2～4 之间，本次试验分别验证 2，2.5，3，3.5 的效果；

2)RANSAC 算法中数据是否符合模型的偏差阈值，选择 0.5～8 之间，分辨率为 0.1。

同时，在传统 SIFT 特征点参数之外，将 SIFT 提取时高斯金字塔的层数增加到16，初始高斯模糊参数选择为 1.3，查看实验效果。最后，针对当前深度学习用于目标检测效果不断提升的现状，通过人工添加目标检测框，模拟位姿估计算法和目标检测算法结合使用时的效果。

3.3　实验结果

在 3 个自由度上的误差如图 3～图 5 所示，其中，X 轴为 RANSAC 的偏差阈值，Y 轴为旋转 10° 时，实际测得的转动角度和 10°的差值。

1)绕 X 轴旋转结果

如图 3 所示，上图为 8 层高斯金字塔效果，下图为 16 层高斯金字塔效果。

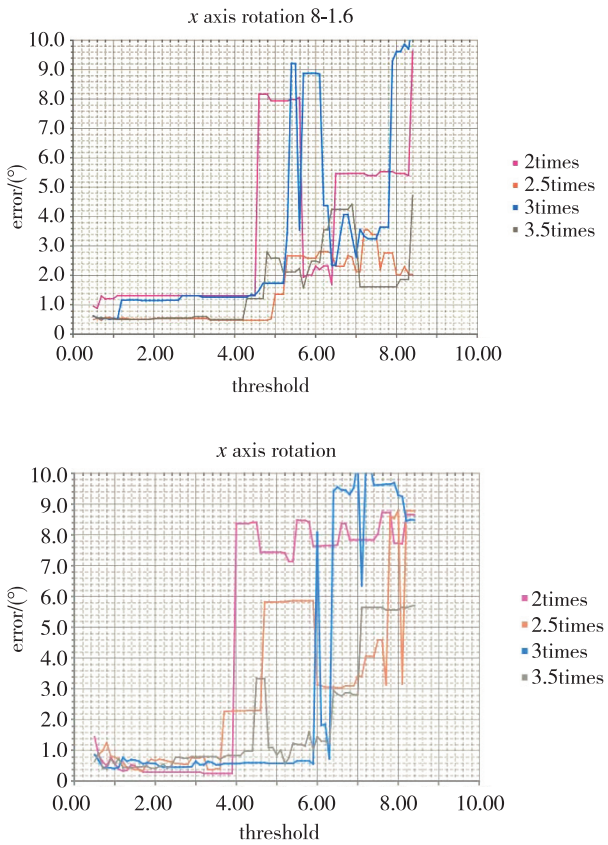

图 3　绕 *X* 轴旋转结果图

Fig. 3　Result of rotate around the *X* axis

2)绕 Y 轴旋转结果：

如图 4 所示，上图为 8 层高斯金字塔效果，下图为 16 层高斯金字塔效果。

3)绕 Z 轴旋转结果：

如图 5 所示，上图为 8 层高斯金字塔效果，下图为 16 层高斯金字塔效果。

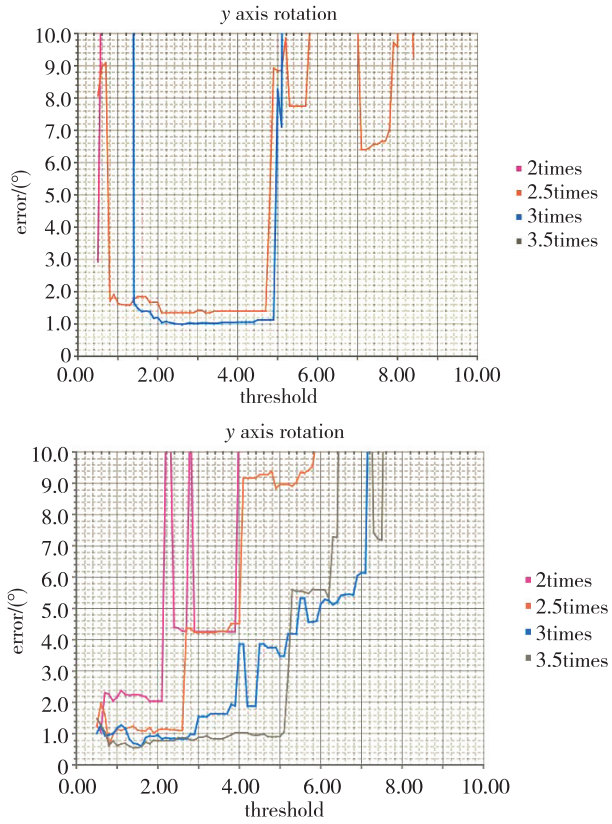

图 4 绕 *Y* 轴旋转效果图

Fig. 4 Result of rotate around the *Y* axis

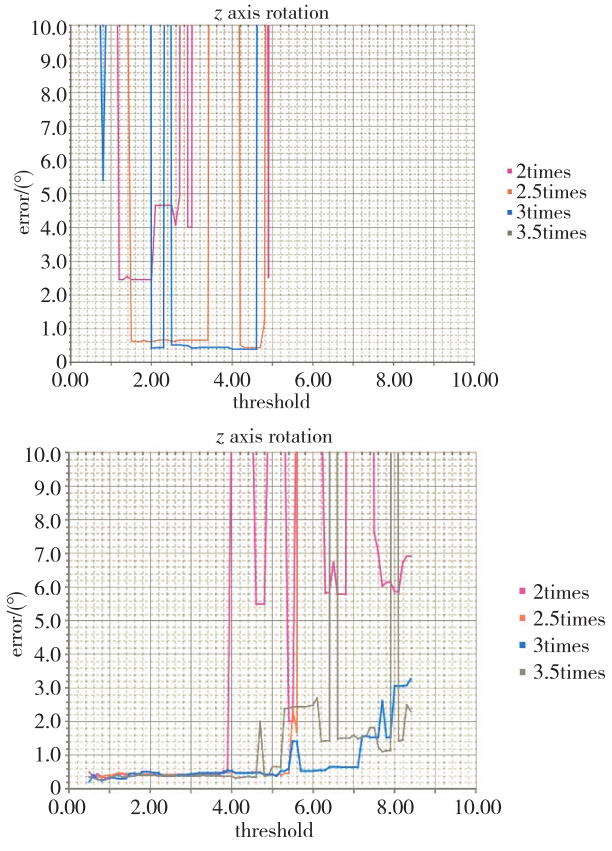

图 5 绕 *Z* 轴旋转效果图

Fig. 5 Result of rotate around the *Z* axis

3.4 实验分析

实验结果表明：

1)RANSAC 偏差阈值越大，允许的误差就越大，不稳定性越高，绝大多数情况下，允许偏差在到达一定值之后，都会使得位姿估计出现错误。

2)输入 PNP 算法中的 SIFT 点对选择倍数 n，过小则被误舍弃的正确匹配点对过多，导致不稳定。

3)提高 SIFT 特征点的高斯金字塔层数可以明显增多 SIFT 特征点的数量，提高算法的稳定性、实时性和精度。

4)偏差阈值选择 2～3 之间，SIFT 点选择倍数在 3～3.5 之间，在标准高斯金字塔的基础上将层数翻倍，可以取得最好的效果。

4 结论

本文提出了一种基于 SIFT 特征匹配和 PNP 算法的单目位姿估计方法，并进行了实验验证和参数确定。结果表明：

1)该算法可以通过选择合适的参数，在稳定性、实时性和精度方面满足机器人抓取的需求，拓宽了机器人的应用范围。

2)根据不同的任务需求，可以在稳定性、实时性和精确度方面进行取舍，本文的实验结果可以作为参考。

3)本文还可以作为点云粗配准的一种方案，配合点云精配准取得更好的结果。

参考文献

[1] 王修岩，程婷婷．基于单目视觉的工业机器人智能抓取研究[J]．机械设计与制造，2011(5)：135-136.
Wang Xiuyan, Cheng Tingting. Rearch on industrial robot intelligent grasp based on monocular vision[J]. Machinery Design & Manufacture, 2011(5): 135-136 (in Chinese).

[2] 逯建军，任晓军，孙伟，等．惯性/双目视觉里程计深组合导航方法[J]．导航定位与授时，2016，3(3)：37-43.
Lu Jianjun, Ren Xiaojun, Sun Wei, et al. INS/Stereo visual odometry deeply integrated navigation method [J]. Navigation Positioning & Timing, 2016, 3(3): 37-43(in Chinese).

[3] 屈桢深，楚翔宇，赵霄洋，等．一种基于改进 Kalman 滤波的视觉/惯性组合导航算法[J]．导航定位与授时，2017，4(2)：14-20.

Qu Zhenshen, Chu Xiangyu, Zhao Xiaoyang, et al. A visual/inertial integrated navigation algorithm based on improved Kalman filter[J]. Navigation Positioning & Timing, 2017, 4(2): 14-20(in Chinese).

[4] 王旒军，陈家斌，余欢，等. RGB-D SLAM 综述[J]. 导航定位与授时，2017，4(6)：9-18.
Wang Yujun, Chen Jiabin, Yu Huan, et al. An overview of RGB-D SLAM[J]. Navigation Positioning & Timing, 2017, 4(6): 9-18(in Chinese).

[5] 张超，王芳，李楠. 基于视觉的惯性导航误差在线修正[J]. 导航定位与授时，2018，5(3)：104-110.
Zhang Chao, Wang Fang, Li Nan. The online correction of IMU biases for visual-inertial navigation[J]. Navigation Positioning & Timing, 2018, 5(3): 104-110(in Chinese).

[6] Zhang Z. A flexible new technique for camera calibration[M]. IEEE Computer Society, 2000.

[7] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[8] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]// International Conference on Computer Vision. IEEE, 2012: 2564-2571.

[9] Bay H, Tuytelaars T, Van Gool L. SURF: Speeded up robust features[C]// European conference on Computer Vision. Springer, Berlin, Heidelberg, 2006: 404-417.

[10] Harris C G, Stephens M. A combined corner and edge detector[C]// Proceedings of 4th Alvey Vision Conference, 1988: 147-151.

[11] Gao X S, Hou X R, Tang J, et al. Complete solution classification for the perspective-three-point problem [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(8): 930-943.

[12] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate O(n) solution to the PnP problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166.

[13] Hesch J A, Roumeliotis S I. A direct least-squares (DLS) method for PnP[C]// 2011 International Conference on Computer Vision. IEEE, 2011: 383-390.

[14] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of the ACM, 1981, 24(6): 381-395.

[15] Hu Z, Uchimura K. UV-disparity: an efficient algorithm for stereovision based scene analysis[C]// IEEE Intelligent Vehicles Symposium Proceedings. IEEE Computer Society, 2005: 48-54.