基于单目 RGB 数据的三维模板物体跟踪算法综述

宋修强[1,2)], 金立[1,2)], 宋婧[1,2)], 李佳宸[3)], 孟祥旭[1,2)], 秦学英[1,2)*]

[1)] (山东大学软件学院 济南 250101)
[2)] (数字媒体技术教育部工程研究中心 济南 250101)
[3)] (浙江大学 CAD&CG 国家重点实验室 杭州 310030)
(qxy@sdu.edu.cn)

摘 要: 文中介绍了三维物体跟踪技术及其最新研究进展, 包括理论基础与评估指标、基于视觉特征的传统位姿求解和基于学习的位姿估计. 分析了传统算法中各种特征约束、统计分割模型、能量函数以及加速策略. 其中, 区域及边缘约束成为跟踪的主流方向, 多种特征可以融合产生更强的约束. 统计分割模型由全局模型逐步发展为局部模型. 非线性最小二乘形式的能量函数易于优化而被广泛采用. 预计算、预渲染的加速策略使得跟踪的实时性极大提高. 基于学习的算法现阶段在精度和速度上不及传统算法, 但展现了更好的特征提取能力和处理更复杂场景的潜力. 梳理了多种算法的跟踪精度及取得高精度的原因. 总结了三维物体跟踪在复杂情境下跟踪精度下降和失败的问题, 及其潜在的解决方案与发展方向. 三维物体跟踪正朝向多特征融合、预计算、多任务的方向发展.

关键词: 三维物体跟踪; 姿态估计; 增强现实
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20155

A Survey of 3D Template Object Tracking Algorithms Based on Monocular RGB Data

Song Xiuqiang[1,2)], Jin Li[1,2)], Song Jing[1,2)], Li Jiachen[3)], Meng Xiangxu[1,2)], and Qin Xueying[1,2)*]

[1)] (*School of Software, Shandong University, Jinan* 250101)
[2)] (*Engineering Research Center of Digital Media Technology, Ministry of Education, Jinan* 250101)
[3)] (*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou* 310030)

Abstract: Introduced the 3D object tracking technology and its latest research progress, including the theoretical foundation and evaluation metrics, traditional pose estimation based on visual features, and learning-based pose estimation. Analyzed various feature constraints, statistical segmentation models, energy functions, and acceleration strategies in traditional algorithms. Region and edge constraints become the mainstream, and fused features can produce stronger constraints. The statistical segmentation model has gradually developed from a global model to local models. The energy function in the form of nonlinear least squares is easy to optimize and is widely used. Pre-computation and pre-rendering acceleration strategies greatly improve the real-time performance. Learning-based algorithms are not as accurate and fast as traditional algorithms, but show greater potential of feature extraction and handling more complex scenes. Summarized tracking accuracy of various algorithms and the reasons for high accuracy. Reviewed the problems

收稿日期: 2023-02-15; 修回日期: 2023-09-15. **基金项目**: 国家重点研发计划(2020YFB1708903); 国家自然科学基金(62172260). **宋修强**(1997—), 男, 博士研究生, CCF 会员, 主要研究方向为三维物体跟踪; **金立**(1998—), 男, 博士研究生, 主要研究方向为类级别三维物体检测; **宋婧**(1997—), 女, 硕士, 主要研究方向为三维物体跟踪; **李佳宸**(1993—), 男, 博士, 主要研究方向为三维物体跟踪; **孟祥旭**(1962—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为人机交互与虚拟现实、计算机图形学、CAD/CAM/CIMS、科学计算可视化; **秦学英**(1966—), 女, 博士, 教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为增强现实、计算机视觉、计算机图形学.

of tracking accuracy decline and failure in complex situations, as well as potential solutions and development directions. 3D object tracking is moving towards the direction of multi-feature fusion, pre-computation, and multitasking.

Key words: 3D object tracking; pose estimation; augmented reality

三维物体跟踪是计算机视觉中的一项关键技术. 给定刚性三维物体模型, 当前帧图像及初始位姿, 三维物体跟踪实时估计相机与三维物体之间的相对位置与姿态关系, 即物体位姿, 亦称相机外参, 包括 3 自由度旋转参数和 3 自由度平移参数. 三维物体跟踪是视觉低层语义和高层语义之间的桥梁, 它能够利用底层的特征估计位姿, 并作为高层语义分析的基础. 目前, 三维物体跟踪在数字孪生、装配引导、AR 游戏、机器人抓取等领域都有重要的应用[1-2].

三维物体跟踪中的物体模型通常通过 CAD 建模获得, 在实际场景中也可以通过预先重建得到. 三维模型作为跟踪过程中的模板, 使物体起到自然三维标记的作用, 且不会对环境产生侵入式的干扰. 三维模型能够为三维物体跟踪提供轮廓、深度等的先验, 并为无纹理或弱纹理的三维物体跟踪提供重要的约束. 初始位姿可以通过位姿估计算法[3]等获得, 在实际场景中也可以手动设置物体位姿. 由于单目彩色相机的廉价性及广泛应用, 因此基于单目 RGB 图像的三维物体跟踪算法成为跟踪算法研究的热点. 本文将重点介绍基于单目 RGB 视频的三维物体跟踪算法.

三维物体跟踪算法根据优化位姿时的约束条件可以分为基于特征点的算法、直接法、基于边缘的算法、基于区域的算法、基于显式特征融合的算法以及基于学习的算法. 本文首先对三维物体跟踪的理论基础进行介绍, 然后分别介绍相关算法.

1 三维物体跟踪理论基础

本节将介绍三维物体跟踪中的理论基础, 包括相机模型、位姿参数化与位姿求解.

1.1 相机模型

对于世界坐标系下的三维模型点 $\boldsymbol{X}=[X_{\mathrm{w}},Y_{\mathrm{w}},Z_{\mathrm{w}}]^{\mathrm{T}}$, 其对应的齐次坐标为 $\tilde{\boldsymbol{X}}=[X_{\mathrm{w}},Y_{\mathrm{w}},Z_{\mathrm{w}},1]^{\mathrm{T}}$. 使用针孔相机模型对其投影可以得到该点的像素坐标 $\boldsymbol{u}=[u_{\mathrm{p}},v_{\mathrm{p}}]^{\mathrm{T}}$ 和齐次坐标 $\tilde{\boldsymbol{u}}=[u_{\mathrm{p}},v_{\mathrm{p}},1]^{\mathrm{T}}$, 投影变换可以表示为 $\boldsymbol{u}=\psi(\boldsymbol{K}(\boldsymbol{T}\tilde{\boldsymbol{X}})_{3\times1})$. 其中, 函数 $\psi(\boldsymbol{X})=[X_{\mathrm{w}}/Z_{\mathrm{w}},Y_{\mathrm{w}}/Z_{\mathrm{w}}]^{\mathrm{T}}$ 能够将相机坐标转换至图像坐标; $\boldsymbol{K}\in\mathbb{R}^{3\times3}$ 表示相机内参, 可以预先标定; $(\cdot)_{3\times1}$ 表示取矩阵的前 3 行第 1 列; $\boldsymbol{T}=\begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\\ \boldsymbol{0}^{\mathrm{T}} & 1\end{bmatrix}\in$ SE(3) 表示物体位姿, 其将世界坐标系映射到相机坐标系. 其中, $\boldsymbol{R}\in$SO(3) 表示旋转矩阵; $\boldsymbol{t}=[t_x,t_y,t_z]^{\mathrm{T}}\in\mathbb{R}^{3\times1}$ 表示平移向量; SO(3) 表示特殊正交群; SE(3) 表示特殊欧几里得群.

1.2 位姿参数化

刚性三维物体跟踪的位姿可以有多种表示方式, 其中平移多用平移向量 $\boldsymbol{t}$ 表示, 旋转的表示方式则有欧拉角、李代数和四元数等. 李代数具有求导的简洁性, 常被用于位姿的表示.

SE(3) 对应的李代数 $\mathfrak{se}(3)$ 为定义在 $\mathbb{R}^{6\times1}$ 上的向量, 即 $\mathfrak{se}(3)=\left\{\boldsymbol{\xi}=\begin{bmatrix}\boldsymbol{\omega}\\ \boldsymbol{\upsilon}\end{bmatrix}\in\mathbb{R}^{6\times1},\hat{\boldsymbol{\xi}}\in\mathbb{R}^{4\times4}\right\}$. 其中, $\hat{\boldsymbol{\xi}}$ 为李代数 $\boldsymbol{\xi}$ 对应的反对称矩阵. 李代数 $\boldsymbol{\xi}$ 的具体形式为 $\boldsymbol{\xi}=\begin{bmatrix}\boldsymbol{\omega}\\ \boldsymbol{\upsilon}\end{bmatrix}=[\omega_1,\omega_2,\omega_3,\upsilon_1,\upsilon_2,\upsilon_3]^{\mathrm{T}}$. 其中, 三维向量 $\boldsymbol{\omega}=[\omega_1,\omega_2,\omega_3]^{\mathrm{T}}\in\mathfrak{so}(3)$ 是 SO(3) 对应的李代数, 用于表示旋转; 三维向量 $\boldsymbol{\upsilon}=[\upsilon_1,\upsilon_2,\upsilon_3]^{\mathrm{T}}$ 用于表示平移. 位姿 $\boldsymbol{T}$ 通过对数映射 ln(·) 可以得到 $\boldsymbol{\xi}$ 的反对称矩阵 $\hat{\boldsymbol{\xi}}=\ln(\boldsymbol{T})$. 通过对 $\boldsymbol{\xi}$ 的反对称矩阵 $\hat{\boldsymbol{\xi}}$ 进行指数映射 exp(·) 可以获得物体位姿 $\boldsymbol{T}=\exp(\hat{\boldsymbol{\xi}})$. 关于李代数详细的推导及数学证明可以参考文献[4-6].

1.3 位姿求解

三维物体跟踪的主要流程如图 1 所示, 给定物体的三维 CAD 模型、初始位姿及当前帧图像, 三维物体跟踪通过优化初始位姿得到当前帧位姿.

传统的三维物体跟踪的位姿求解可以等价为求解当前帧与参考帧(通常情况下为上一帧)之间的对应关系. 基于特征点匹配的算法求解特征点之间的对应关系, 基于边缘的算法通常求取前景边缘点与投影轮廓点之间的对应关系, 基于区域的算法求解前景投影轮廓与当前帧最佳分割之间的

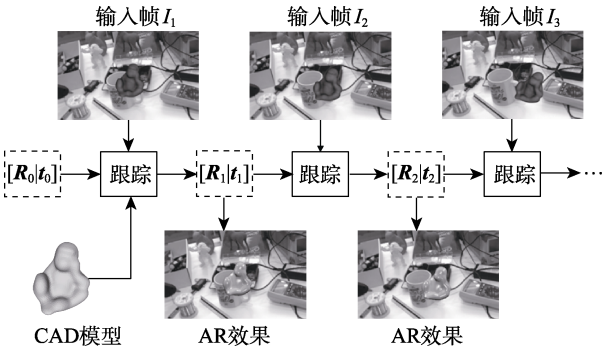


图 1　三维物体跟踪的主要流程

对应关系，直接法则求解前景点之间的对应关系. 在李代数的表示下，帧间位姿增量的能量函数为

$$\Delta\boldsymbol{\xi}_t = \arg\min_{\Delta\boldsymbol{\xi}_t} E(\Delta\boldsymbol{\xi}_t) = \arg\min_{\Delta\xi_t}(\boldsymbol{P}_{t-1}, \boldsymbol{P}_t, \boldsymbol{\xi}_{t-1}, \Delta\boldsymbol{\xi}_t).$$

其中，$\boldsymbol{P}_{t-1}$ 为上一帧中提取的点或三维模型的投影点；$\boldsymbol{P}_t$ 为当前帧中提取的点或待求解点；$\boldsymbol{\xi}_{t-1}$ 为上一帧位姿(初始位姿)；$\Delta\boldsymbol{\xi}_t$ 为当前帧相对于上一帧的位姿增量. 三维物体跟踪是一个非线性优化问题，可以通过随机梯度下降法、L-M 算法以及高斯-牛顿法等对位姿增量进行求解. 求得位姿增量后，对位姿进行更新 $\boldsymbol{T} \leftarrow \exp(\Delta\hat{\boldsymbol{\xi}}_t)\boldsymbol{T}$，即完成对当前帧的跟踪. 基于跟踪结果可以完成一些增强现实应用.

2　三维物体跟踪的评估方法

三维物体跟踪算法的重要评估指标是速度和精度，一般要求算法速度能够达到 30 FPS，对于有强实时要求的场景，要求算法能够达到 60 FPS. 目前有多个数据集及评估指标用于评估三维物体跟踪算法. 本节将对三维物体跟踪算法常用的数据集及评估指标进行介绍.

2.1　数据集

2.1.1　RDOT 数据集

RDOT 数据集[7]是一个合成数据集，包含 6 个场景和 27 个物体，并在正常图像的基础上增加了光照变换、噪声、运动模糊和遮挡. 视频分辨率为 720×480 像素，共计 1 603 800 帧图像. 该数据集的优点是物体数量丰富，并且存在半透明、高光材质的物体；此外，该数据集场景丰富，包括室内场景和室外场景. 缺点是该数据集是合成数据集，完全由渲染得到.

2.1.2　Choi 数据集

Choi 数据集[8]是一个点云数据集，包含 4 个物体，4 个渲染序列，共 4 000 帧图像，图像分辨率为 640×480 像素. 该数据集的优点是它为 RGB 图像提供了准确的点云数据；缺点是物体数量较少、运动模式较为单一、图像序列较少，且背景较为简单.

2.1.3　RBOT 数据集

RBOT 数据集[9]是一个半合成数据集，包含 18 个物体，4 个场景(常规、动态光照、噪声以及遮挡)，共计 72 个序列，每个序列有 1 000 帧用于测试的图像，图像分辨率为 640×512 像素. 该数据集的背景为真实场景，通过渲染得到前景物体，物体的运动速度较快. 它是当前三维物体跟踪最常用的测试数据集，其优点是它包含了较多数量形状、纹理各异的物体，视频序列中存在复杂的背景，并且每个视频序列存在 4 种变换，包含具有挑战性的场景. 其缺点是它的前景物体由渲染得到，和真实环境的光照并不一致；同时各个物体的运动轨迹单一，所有物体共享一组运动轨迹；此外，该数据集的场景也比较单一.

2.1.4　OPT 数据集

OPT 数据集[10]是一个真实的数据集，包含 6 个物体，552 个用于三维物体跟踪测试的视频序列，视频分辨率为 1 920×1 080 像素，共有 79 968 帧图像；它们是在不同的光照环境下，利用机械臂手持相机以不同运动路径、不同速度拍摄的. 该数据集的优点是其数据由实拍获得；缺点是背景简单，并且运动路径受限于机械臂的活动范围.

2.1.5　YCB-Video 数据集

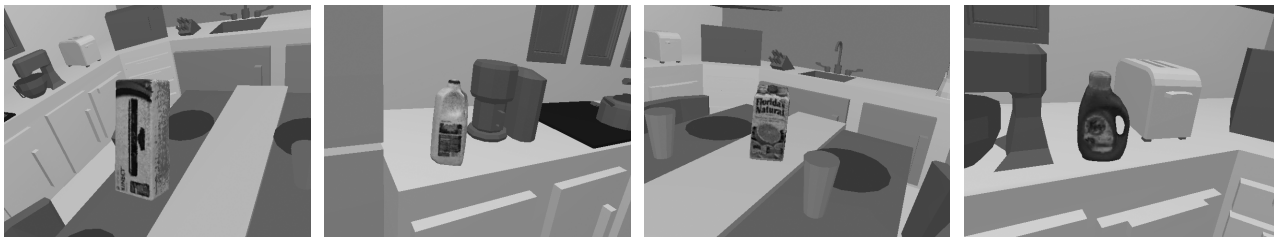
YCB-Video 数据集[11]是一个真实的数据集，包含 21 个物体，92 个视频序列，图像分辨率为 640×480 像素. 物体在视频中保持与场景的静止，并有一些自然噪声. 该数据集中的视频序列由 RGB-D 相机在室内拍摄得到，场景由 21 个物体中的 3~9 个物体搭建获得. 该数据集的优点是数据由实拍获得，并且物体以及场景多样化. 缺点是真实位姿的标注误差较大.

2.1.6　BCOT 数据集

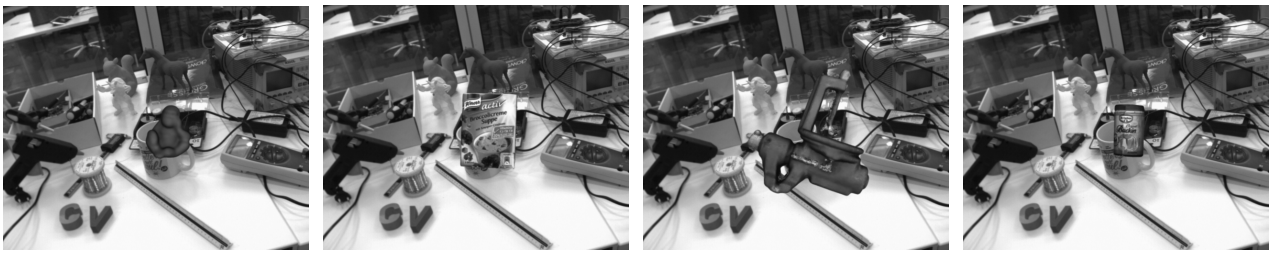
BCOT 数据集[12]是一个实景拍摄的数据集，包含 20 个物体及 22 个场景，图像总数量达 126 000 帧，包括室内场景和室外场景，其中一部分数据为双目数据，图像分辨率为 640×512 像素. 物体的真实位姿通过双目跟踪算法[12]获得，标注的平移误差在 2 mm 以内. 该数据集的优点是数据由实拍获得，且物体和场景多样化，很好地反映了现实世界的场景；同时，该数据集包含有光照变换和遮挡，且含有双目数据；此外，物体的运动轨迹多样，在不同视频序列中，物体相对于相机的运动均不同. 缺点是相对于渲染数据集而言，位姿真实值略有误差. 上述 6 个数据集的部分场景如图 2 所示.



a. RDOT 数据集



b. Choi 数据集



c. RBOT 数据集



d. OPT 数据集



e. YCB-Video 数据集



f. BCOT 数据集

图 2　三维物体跟踪常用数据集部分示例展示

2.2　评估指标

2.2.1　*n* cm/*n*°

n cm/*n*°是三维物体跟踪中最常用的评估指标. 当前帧跟踪结果平移误差不超过 *n* cm, 且旋转误差不超过 *n*°即认为跟踪成功, 否则认为跟踪失败, 然后读取位姿真值并重新初始化. 跟踪结束后统计跟踪成功的帧所占的比例, 即为跟踪精度. 最常用的指标为 5 cm/5°, 此外还有 2 cm/2°, 1 cm/10°等.

平移误差 e_{t} 和旋转误差 e_{r} 的计算方式分别为

$$e_{\mathrm{t}} = \| \hat{\boldsymbol{t}} - \boldsymbol{t} \|_2 ,$$

$$e_{\mathrm{r}} = \arccos\left(\frac{\mathrm{trace}\left(\hat{\boldsymbol{R}}^{\mathrm{T}} \boldsymbol{R} \right) - 1}{2} \right).$$

其中, $\boldsymbol{R}$ 和 $\boldsymbol{t}$ 为真实位姿; $\hat{\boldsymbol{R}}$ 和 $\hat{\boldsymbol{t}}$ 表示估计值.

2.2.2　平均距离

平均距离(average distance, ADD)指标评估模

型上的点在估计的位姿和真实位姿下的平均距离误差，即

$$\mathrm{ADD}=\frac{1}{m}\sum_{\boldsymbol{X}\in M}\|(\boldsymbol{R}\boldsymbol{X}+\boldsymbol{t})-(\hat{\boldsymbol{R}}\boldsymbol{X}+\hat{\boldsymbol{t}})\|_2.$$

其中，M 为三维模型上所有点的集合；m 为三维模型上所有点的数量；$\boldsymbol{X}$ 为三维模型上的点.

2.2.3　平均最近点距离

平均最近点距离(average closet point distance, ADD-S)指标多用于评估对称物体，其评估三维模型上的点在真实位姿下与对应的最近点之间的平均距离，即

$$\mathrm{ADD-S}=\frac{1}{m}\sum_{\boldsymbol{X}_1\in M}\min_{\boldsymbol{X}_2\in M}\left\|\left(\boldsymbol{R}\boldsymbol{X}_1+\boldsymbol{t}\right)-\left(\hat{\boldsymbol{R}}\boldsymbol{X}_2+\hat{\boldsymbol{t}}\right)\right\|_2.$$

其中，$\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 为三维模型上的点.

3　三维物体跟踪算法

如前所述，三维物体跟踪算法可以分为基于特征点的算法、直接法、基于边缘的算法、基于区域的算法、基于显示特征融合的算法，以及基于学习的算法. 近年来，多种基于边缘和区域的算法在精度和速度方面都取得了很好的效果，是较为主流的算法. 基于特征融合的算法使用多种特征提供对位姿的约束，也取得了很高的跟踪精度. 基于学习的算法近年来逐渐成为三维物体跟踪的一个重要方向，但其精度和速度尚不及传统算法.

3.1　基于特征点的算法

基于特征点的三维物体跟踪算法以显著特征点作为基础，建立二维图像点和三维空间点的对应关系，并通过求解此关系解算物体的位姿. 常用的求解方法有直线线性变换[13]，EPnP(efficient perspective-n-point)[14] 和 UPnP(unified perspective-n-point)[15]等.

早期的一些基于特征点的算法不使用三维模型，而是以预先标定位姿的多幅二维图像作为先验. Kato 等[16]在物体上预先设置标记物，使得标记物与待跟踪物体之间保持相对静止，在图像中检测标记物的顶点，并根据标记物特殊的编码特征确定二维点与空间三维点的对应关系. 因为需要设置标记物，并且标记物需要始终出现在视野中，所以会对环境产生侵入，具有很大的局限性. 之后，特征点检测和特征描述技术的不断发展[17-19]，使得不借助标记物，只借助物体上鲜明的纹理提供的关键点进行跟踪成为可能[20-23]. Neumann 等[23]使用检测得到图像特征点的描述子，对 2 帧图像之间的特征点进行匹配，并通过随机抽样一致性去除外点，进而计算 2 帧之间物体相对相机位姿的增量，计算当前帧物体的位姿.

以三维模型作为先验的算法[24-25]使用待跟踪物体的三维模型上的固定特征，将其投影到二维图像上，以消除长序列跟踪带来的漂移问题；通过最小二乘拟合得到跟踪物体的位姿，但优化位姿时有可能陷入局部极值. 离线相机配准的发展，使得能够对图像之间的特征点进行全局的集束约束，可以获得高精度的地图定位，这给三维跟踪带来了新的思路，Vacchetti 等[26]使用集束约束选取关键帧的方法消除跟踪过程中出现的漂移和抖动.

同时，有些工作通过改进特征点检测的精度来改善三维物体跟踪的精度[27]，也有一些工作[28]通过改进描述子的计算量提高三维物体跟踪的速度. 基于特征点的算法十分依赖待跟踪物体的纹理特征，因此其难以应对无纹理或弱纹理的物体.

3.2　直接法

直接法基于短时颜色一致性假设[29]，该假设认为短时间内 2 帧图像上对应点的颜色是不变的，因此它可以用于寻找 2 帧图像上点之间的对应. 最简单的位姿求解形式为

$$\Delta\boldsymbol{\xi}=\arg\min_{\Delta\boldsymbol{\xi}}E(\Delta\boldsymbol{\xi})=$$

$$\arg\min_{\Delta\boldsymbol{\xi}}\sum_{i=1}^{n}(I_{t-1}(\boldsymbol{u}_i)-I_t(\boldsymbol{u}_i(\Delta\boldsymbol{\xi})))^2.$$

其中，$I_{t-1}(\boldsymbol{u}_i)$ 为上一帧中像素 $\boldsymbol{u}_i$ 的灰度值；$I_t(\boldsymbol{u}_i(\Delta\boldsymbol{\xi}))$ 为当前帧中待寻找的对应像素处的灰度值；$E(\Delta\boldsymbol{\xi})$ 取得最小值时，即可获得当前帧的位姿. 直接法优化得到的当前帧与上一帧之间点的一组对应如图 3 所示.



a. 上一帧的点　　b. 对应的当前帧上的点

图 3　直接法优化得到的相邻 2 帧间点的一组对应点

早期的一些算法利用灰度值差的平方来计算能量函数，但在实际场景中，图像容易受到光照变

化的影响，导致短时一致性假设不够准确．对于非朗伯表面或镜面反射，该假设也难以成立．因此，一些算法[30-31]为像素的灰度信息添加了额外的约束项，还有一些算法则致力于建立对像素更鲁棒的特征表示．Crivellaro 等[32]提出了描述子场，这是一种稠密的描述子，在每个像素处计算一个局部描述子，是对像素的一种向量描述．与像素的颜色描述相比，该描述子对光照变化更加鲁棒．Seo 等[33]在朗伯假设下，采用物体的表面法向量刻画像素值的变化，并提出了一个受约束的目标函数以减小积累误差．

与基于特征点的算法相比，在前背景具有明显差异时，直接法能够跟踪无纹理或弱纹理的物体．但是直接法容易受到光照变化的影响，会产生积累误差，同时其忽略了前景物体的边缘信息．

3.3 基于边缘的算法

对于无纹理或弱纹理物体，显著的纹理特征难以获取，三维模型的投影轮廓因此成为更为可靠的信息．基于边缘的三维物体跟踪算法是解决无纹理或弱纹理物体跟踪的重要算法之一．此类算法利用上一帧的位姿信息，将三维模型投影至当前帧的图像上，通过最小化三维模型投影轮廓点与图像前景边缘点的差异来优化位姿．

Lowe[34]将图像中的边缘点连为线段，寻找其与三维模型投影轮廓线段的对应，进而通过最小化二者的欧几里得距离求解位姿．该算法的准确度严重受到线段之间对应关系准确率的影响，且只适用于规则的多边形物体．Klein 等[35]提出了基于粒子滤波的跟踪算法，粒子滤波的使用对排除干扰的边缘线段和估计边缘相似度具有积极作用，对复杂形状的三维物体的跟踪更加鲁棒．

Harris 等[36]提出了首个能够达到实时的三维物体跟踪算法，其首先利用初始位姿将模型上的可见三维点投影到二维图像上，然后在投影轮廓点的法向上寻找对应的图像边缘点，通过二维图像点和三维模型点的对应关系，并使用最小二乘法迭代计算出物体的位姿．但当在三维模型投影轮廓附近存在噪声时，算法可能会产生错误的匹配，导致位姿求解的错误．

Wang 等[37]采用构建图模型解决杂乱环境中三维模型投影轮廓和图像中候选边缘之间的配准问题，并利用动态规划寻找能够最大概率分割背景且相邻连续的点，将其作为物体的边缘点．Wang 等[38]提出一种边缘距离场，隐式地寻找投影轮廓点与前景边缘点的对应，直接最小化二者的欧几里得距离求解位姿，并使用粒子滤波预测位姿初值，以避免落入局部极小值．Wang 等[39]提出了一种应对遮挡和快速移动物体的三维跟踪算法，通过三维模型投影轮廓点及图像前景边缘点的法向差异计算边缘一致性分数，并将一致性分数融入到粒子滤波中，以减少外点的影响．Huang 等[40]使用颜色信息过滤三维模型投影轮廓附近的噪声边缘，以更好地区分前景和背景；通过联合优化边缘距离和颜色相似度，解决了杂乱环境和遮挡情形下的实时跟踪问题，并把该算法扩展到存在相互遮挡的多目标跟踪上．Sun 等[41]根据角点检测得分对三维模型投影轮廓进行分割，并在轮廓梯度方向上匹配图像前景边缘，然后使用 P*n*P (perspective-*n*-point)求解物体位姿，该算法有效地优化了在杂乱背景中对物体的跟踪．Dong 等[42]提出了基于数学分析模型的优化算法，并提出了边缘点自适应权重的算法，以解决跟踪过程中的遮挡和边缘不清晰的问题，提高了跟踪的整体表现．

Tian 等[43]使用颜色统计模型计算前景的概率图，利用一种长搜索线结构在概率图上对物体的位移进行搜索，它在帧间大位移的情况下具有明显的优势．同时该文指出三维物体跟踪算法的失败主要源于平面外旋转，因此其对物体的旋转预先进行搜索，然后进行优化．该算法取得了很高的精度，同时它在跟踪前进行了多种预计算，具备很好的实时性．

基于边缘算法的稳定性受图像中提取的边缘质量的影响，在物体纹理丰富或者背景杂乱时，会受到较多无关边缘的影响，可能会造成物体跟踪失准．

3.4 基于区域的算法

基于区域的算法同时进行对当前帧图像前背景的分割及位姿估计，能最优分割前背景的曲线即物体在最优位姿下的投影轮廓．早期的一些算法[44-46]通常是通过对物体显式分割[47]得到其轮廓，然后进行位姿的优化．

图像分割的过程时间开销很大，因此上述算法难以达到实时．后来的一些算法不再进行显式的分割，而是使用像素级后验概率对前背景进行隐式分割．像素的后验概率，即像素属于前景和背景的概率，需要通过统计分割模型来进行计算，因此良好的统计分割模型是该类算法的关键．

Osher 等[48]提出使用像素级后验概率的 PWP3D 算法，对图像中前景和背景中所有像素点的颜色进行统计，并建立全局的颜色直方图，据此

可以计算给定像素点属于前景或背景的后验概率. 同时 PWP3D 使用水平集函数[49]建立物体投影轮廓与前背景区域之间的关系. 水平集函数是闭合曲线的隐式表示，隐式地将一条曲线定义为其零水平集. 对于二维闭合曲线 $\boldsymbol{C}$，有

$$\boldsymbol{C}=\left\{\boldsymbol{u}\in\mathbb{R}^2 \mid \Phi(\boldsymbol{u})=0\right\}.$$

其中，$\boldsymbol{u}$ 表示图像上的像素点；$\Phi(\cdot)$ 表示有向距离场函数，即

$$\Phi(\boldsymbol{u})=\begin{cases}-D(\boldsymbol{u},\boldsymbol{C}), & \forall \boldsymbol{u}_i\in\Omega_{\mathrm{f}}\\ D(\boldsymbol{u},\boldsymbol{C}), & \forall \boldsymbol{u}_i\in\Omega_{\mathrm{b}}\end{cases}.$$

其中，$D(\boldsymbol{u},\boldsymbol{C})=\min\limits_{\boldsymbol{u}_i\in\boldsymbol{C}}\|\boldsymbol{u}-\boldsymbol{u}_i\|_2$ 表示点 $\boldsymbol{u}$ 与距离其最近的轮廓点的欧几里得距离；Ω_{f} 表示前景区域；Ω_{b} 表示背景区域. 通过能量函数可以将采样像素点与其前背景概率进行统一，能量函数为

$$E(\Phi)=-\sum_{\boldsymbol{u}_i\in\Omega}\ln\Big(H\big(\Phi\big(\boldsymbol{u}_i(\xi)\big)\big)P_{\mathrm{f}}(\boldsymbol{u}_i)+\Big(1-H\big(\Phi\big(\boldsymbol{u}_i(\xi)\big)\big)\Big)P_{\mathrm{b}}(\boldsymbol{u}_i)\Big).$$

其中，$H(\cdot)$ 表示光滑函数；P_{f} 和 P_{b} 分别表示像素点属于前景和背景的后验概率. PWP3D 使用梯度下降法来对位姿进行迭代求解，然后再更新统计分割模型. PWP3D 利用 GPU 进行了加速，是首个达到实时的基于区域的三维物体跟踪算法. 但是，PWP3D 使用的全局统计分割模型难以刻画图像前景边缘附近的精细信息，且在全局对颜色进行统计会向直方图中加入许多价值不高的信息，造成时间上的浪费. 此外，其使用的梯度下降求解算法收敛精度不高.

如图 4 所示，Tjaden 等[9,50-51]采用高斯-牛顿法来求解物体位姿，其收敛性更好，且对旋转具有更好的鲁棒性. Tjaden 等[9,51]对三维模型点进行投影，并对以投影轮廓点为圆心、半径为 r 的圆形区域内的像素颜色值进行统计，建立颜色直方图，计算统计分割模型.

颜色直方图的数量往往可达数千个，但是出于实时性的考量，在进行一次迭代后，只有一部分颜色直方图会得到更新. 此外，Tjaden 等[9]将能量函数重写为非线性最小二乘的形式，使用高斯-牛顿法进行迭代具有很好的收敛性；并且其在进行位姿优化时采用由粗糙到精细的策略建立图像金字塔，在图像金字塔上逐层进行迭代优化. 即首先在 1/4 分辨率的图像上迭代优化 4 次，然后在 1/2 分辨率的图像上迭代优化 2 次，最后在原图像上迭代优化 1 次. Tjaden 等[9]算法取得了很好的跟踪效果，但是对相似颜色、噪声和遮挡等场景仍然不够鲁棒；同时其颜色直方图也存在更新不及时的问题.

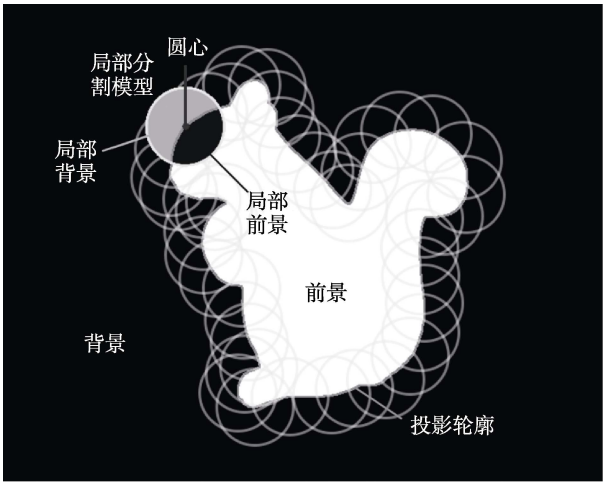


图 4　局部统计分割模型[9]

在极坐标的表示下，Zhong 等[52]对三维模型投影轮廓周边数个局部扇形区域内的像素颜色值进行统计，建立了局部概率分割模型. 与 Tjaden 等[9]算法中的统计分割模型相比，这种扇形统计分割模型的个数大大减少，但是由于每个扇形分割模型需要统计的像素点的数量多，其计算速度并没有得到显著的提升. 其所建立的扇形统计分割模型如图 5 所示.

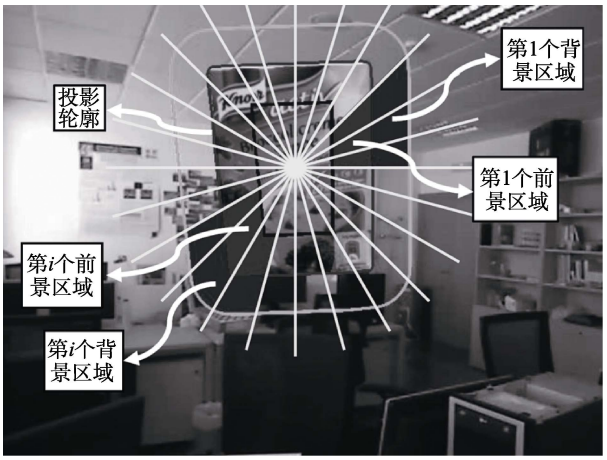


图 5　扇形统计分割模型[52]

扇形统计分割模型的一个优点是其中每个统计分割模型都可以得到及时更新. 此外，Zhong 等[52]采用了边缘距离场[38]的策略来处理遮挡，与 Tjaden 等[9]算法相比，其对遮挡具有更好的鲁棒性.

Huang 等[53]采用了一种搜索线结构，在三维投影轮廓点上沿法线方向建立搜索线. 在搜索线上采样投影轮廓附近的像素点，并利用采样点颜色的后验概率对能量函数进行逐像素加权，能够

完成像素权值的自适应调整. 该算法对遮挡、相似颜色具有较好的鲁棒性.

Stoiber 等[54-55]提出的 RBGT[54], SRT3D[55]算法建立全局的统计分割模型. 这 2 种算法提出了一种对应线的结构, 在三维物体投影轮廓点上建立沿法向的对应线, 利用后验概率来计算对应线上点的前背景概率, 并基于对应线结构进行位姿的优化求解, 其所建立的不同尺度的对应线如图 6 所示.



图 6 RBGT[54]和 SRT3D[55]算法中的对应线

为了加速计算, RBGT 和 SRT3D 算法将对应线上相邻的点组成片段, 并映射至尺度空间, 再以片段为单位进行计算. 三维物体跟踪实时性的一个主要瓶颈是对三维物体的渲染及渲染结果在 GPU 和 CPU 之间的传输. RBGT 和 SRT3D 算法采用了一种预渲染的方式, 在跟踪开始之前, 从不同的视角渲染数千个三维物体的模板; 在跟踪过程中根据当前位姿读取最接近的模板以进行后续优化. 这种方式极大地提高了三维物体跟踪的速度, 这 2 种算法在单核 CPU 上可以达到 1 000 FPS. RBGT 和 SRT3D 算法另一个重要的改进是改进了光滑函数, 使得能量函数符合高斯分布, 因此使用高斯-牛顿法迭代可以在一步之内收敛. 同时, 高斯-牛顿法中增加了吉洪诺夫正则项, 对能量函数的优化更加稳定. RBGT 和 SRT3D 算法每处理一帧图像迭代 14 次, 未使用图像金字塔的迭代方式.

Wang 等[56]延续了 SRT3D 的优化策略, 并进一步处理了位姿快速变化的情形. 其改进了 SRT3D 的模板索引方式, 使用球面网格索引的算法对初始位姿进行搜索; 然后基于该初始位姿实时渲染模板并优化该位姿, 以提升最终的跟踪精度. Stoiber 等[57]提出的 ICG 算法, 进一步利用深度信息将深度约束加入至能量函数, 取得了更好的跟踪效果. 由于预渲染的策略在进行位姿优化时不再对模型进行渲染, 无法找到相邻帧上局部投影点对应的唯一的三维模型点, 因此无法对局部分割概率模型进行更新, 需要使用全局分割概率模型. 这是预渲染策略的一个弊端.

基于区域的算法对杂乱背景、快速运动等场景具有较好的鲁棒性, 但当前背景颜色相似时, 容易跟踪失败. 在图像含有较多噪声, 如在 RBOT 数据集的噪声场景下, 基于区域的算法跟踪精度会有一定程度的降低.

3.5 基于显式特征融合的算法

一些三维物体跟踪算法使用多种特征来进行约束, 并在能量函数中进行了显式地融合. 如使用了区域约束和光度约束的算法[58]、使用区域和特征点约束的算法[59]以及使用区域约束和边缘约束的算法[60-61]等. 与基于单种特征约束的跟踪算法相比, 使用多种特征约束可以提升跟踪算法的鲁棒性.

Zhong 等[58]结合了基于区域的算法和直接法, 对前景边缘附近局部带状区域内的像素颜色进行统计, 并计算统计分割模型, 该模型的数量较少, 一般为几个或十几个. 利用统计分割模型可以为位姿求解提供区域约束, 并使用梯度描述子场[32]提供光度约束. 其能量函数为基于区域的能量函数和直接法的能量函数的加和, 即

$$E = \lambda_1 E_{\text{region}} + \lambda_2 E_{\text{direct}} \tag{1}$$

其中, λ_1 和 λ_2 表示权重系数; E_{region} 表示区域约束下的能量函数; E_{direct} 表示直接法约束下的能量函数. 该算法获得了更加稳定的跟踪效果, 但是对跟踪精度的提升有限.

Li 等[60]同时利用了边缘约束和区域约束, 将采样像素点的区域划分为多个子区域, 分别计算子区域中边缘约束和区域约束的权重, 能够使得 2 类约束起到互补的作用. 其能量函数的形式与式(1)类似.

基于多特征的算法往往需要对一些超参数进行手动设定, 以达到较好的跟踪效果, 这是该类算法的一个缺点.

3.6 基于学习的算法

基于学习的算法采用数据驱动的方式, 建立输入图像和位姿之间的映射关系, 达到三维物体

跟踪的目的. Tan 等[62]提出了基于随机森林的模板跟踪的算法，使用预先密集采样多视角下的三维物体的图像数据对随机森林进行训练. 这种策略对遮挡有更好的鲁棒性，但跟踪的速度达不到实时性的要求. Tan 等[63]提出了联合学习器和优化器的随机森林的算法，提高了基于学习的算法的实时性. 但这 2 种算法都是基于 RGB-D 数据的跟踪算法，在使用场景上受到限制.

随着深度学习技术的发展，其在图像方向上的应用日渐成熟，使得基于图像获得三维信息的能力得到增强. Manhardt 等[64]提出了基于深度学习跟踪的策略，在输入图像和三维物体投影所得图像中，通过网络提取低级和高级特征，并估计 2 帧之间的相对位姿，以完成三维物体跟踪. Li 等[65]提出了 DeepIM 网络，利用初始位姿将三维模型渲染成二维掩模以及对应的彩色前景图，然后与当前帧进行匹配来计算位姿. 为避免前景过小而难以提取有用的特征，该算法首先分割出当前帧图像中前景所在的矩形区域，然后将此区域放大. 该算法同时提出一种迭代网络，循环迭代进行位姿估计以不断减小当前帧的位姿误差. 然而，大量卷积的使用，导致该算法实时性很差，这在多物体位姿估计时尤其明显. Iwase 等[66]借助传统算法优化解决实时性的问题，并提出了一种对三维物体纹理编码的方式，使用网络预测输入图像和三维物体纹理之间的对应关系，进而用高斯-牛顿法优化得到相对位姿. Xu 等[67]提出了一种编码方式应对遮挡，提高了算法的鲁棒性.

基于学习的算法具有更强的特征提取能力，对复杂场景例如噪声场景也展现了更好的鲁棒性. 但是，基于学习的算法需要大量的数据驱动，使得其泛化性能不及传统算法；同时由于网络层数较深，因此通常很难达到实时性的要求.

4　其他相关算法

除上述三维物体跟踪算法之外，还有一些与之类似的算法都能够对三维物体的位姿进行估计.

4.1　类级别 6D 位姿估计

Wang 等[68]创建了标准物体坐标空间(normalized object coordinate space, NOCS)，定义了类级别 6D 位姿和尺度估计问题，并提供了基于 NOCS 的类级别位姿估计数据集，可对类级别未见过的实例物体进行位姿估计. Tian 等[69]提出了类先验模型的概念，并提出了“类先验+形变场+对应矩阵”的位姿估计框架，成为之后类级别位姿估计任务的主流框架. 后来的许多工作，如 Chen 等[70]和 Lin 等[71]都基于文献[69]进行了改进，同时进一步提出了提升类级别位姿估计任务的各项指标. Wang 等[72]通过跟踪 RGB-D 视频中的关键点，提出了一种基于 6D 位姿锚点的类别级关键点跟踪器. Weng 等[73]通过对点云进行规范化，分别对尺度、旋转和位移进行直接回归来估计帧与帧之间的相对位姿. Wen 等[74]通过关键点匹配来估计帧之间的粗略位姿，然后通过位姿图优化和关键帧内存池机制来得到精确位姿. 类级别位姿估计在单帧位姿估计任务上取得了一定的发展，然而对于许多在时间上连续的机器人应用，如即时反馈控制系统和增强现实应用，连续的位姿跟踪会更有帮助.

4.2　基于单帧的位姿估计

基于单帧的位姿估计不需要初始位姿的先验，其通常需要在图像中进行全局搜索，因此实时性不如三维物体跟踪算法. 早期的单帧位姿估计主要基于模板匹配，如 LineMod[75-77]算法. 目前的主流算法为基于深度学习的算法，Xiang 等[11]提出的 PoseCNN，Peng 等[3]提出的 PVNet，以及 Su 等[78]，Xu 等[79]提出的算法，其位姿估计精度较之传统算法有较大的提升.

三维物体跟踪算法估计的是连续帧中物体的位姿，而基于单帧的位姿估计可以估计离散帧中物体的位姿，不需要初始位姿作为先验. 对于三维物体跟踪而言，初始位姿是非常强大的先验，它能够大大缩小位姿搜索的空间，有利于物体位姿的优化. 因此在连续帧的位姿估计中，三维跟踪算法的精度和实时性要超过基于单帧的位姿估计算法. 与此同时，基于单帧的位姿估计可以帮助三维物体跟踪算法从错误的位姿中恢复，这是两者结合的一个重要应用.

5　三维物体跟踪算法在 RBOT 数据集上的跟踪结果

RBOT 数据集自提出以来，已成为三维物体跟踪算法最通用的测试数据集. RBOT 数据集中形状、纹理各异的物体，以及 4 种不同的场景可以涵盖大部分跟踪场景，因此在 RBOT 数据集上的测试结果具有很重要的参考意义. 近年来，许多跟踪精度较高的算法都在 RBOT 数据集上进行了测试，部分算法在该数据集上的测试结果如表 1 所示，表中所用的精度评估指标为 5 cm/5°.

表 1 RBOT 数据集上各算法在不同场景下的跟踪精度对比 %

算法	场景				平均精度
	常规	动态光照	噪声	遮挡	
文献[51]	67.0	67.0	40.6	57.5	58.0
文献[9]	79.9	81.2	56.6	73.3	72.8
文献[58]	80.8	79.0	61.6	74.1	73.9
文献[52]	82.7	81.3	63.6	78.4	76.5
文献[40]	86.9	87.3	65.0	83.6	80.7
文献[54]	90.0	90.6	71.5	85.6	84.4
文献[61]	77.0	79.3	59.2	71.2	71.7
文献[60]	85.8	86.7	71.4	80.3	81.1
文献[41]	88.1	88.8	80.5	85.1	85.6
文献[53]	89.9	90.7	69.6	88.9	84.8
文献[59]	85.5	84.1	63.5	77.6	77.7
文献[55]	94.2	94.6	81.7	93.2	90.9
文献[56]	94.6	95.2	83.1	93.6	91.6
文献[43]	95.2	95.4	83.2	94.9	92.2

SRT3D[54]算法得益于其所提出的光滑函数，取得了很高的跟踪精度. 该光滑函数能够使得能量函数符合高斯分布. Huang 等[53]利用边缘信息对能量函数加以约束，也取得了很好的跟踪效果. Tian 等[43]分析了三维物体跟踪失败的情况大多来自平面外旋转，因此对旋转进行了预搜索，进而执行位姿优化，也取得了很高的跟踪精度.

6 总结与展望

三维物体跟踪近年来在实时性和准确性上取得了较大的提高，并应用于越来越多的实际场景中. 其中，基于边缘和区域的算法逐渐成为近年来的研究热点，能量函数由梯度下降法进行优化发展至由高斯-牛顿法进行优化；利用光滑函数使得能量函数符合高斯分布，有利于优化的进行. 统计分割模型由全局模型逐渐发展为局部分割模型，局部模型具有更好的统计结果. 预渲染、预计算等技术使得三维物体跟踪的实时性产生了巨大的飞跃，但是预渲染技术的应用需要使用全局模型，是其弊端之一. 基于边缘的算法也取得了长足的进展，Tian 等[43]结合了统计分割模型，并专门对平面外旋转进行了搜索，取得了很高的跟踪精度. 多种特征或约束进行结合是提高跟踪精度的一个重要策略，如区域特征和边缘特征的结合，基于区域的算法和直接法的结合等.

基于学习的三维物体跟踪在实时性和准确性上还无法媲美传统算法，且训练耗时，泛化性相比传统算法也较差，这些是需要解决的问题. 由于深度学习技术强大的特征提取能力，因此其有很大的发展空间，特别是在有噪声的场景下，基于学习的算法具有较大的优势. 基于学习的方法与传统方法相结合进行优势互补，是未来的一个发展趋势.

三维物体跟踪具有确定动态物体的位姿变化的能力，在三维模板模型的约束下，通常对于无纹理或弱纹理物体具有快速准确的跟踪能力，这对于在工业场景中孪生被操作对象的运动变化提供了工具，是物体数字孪生的重要基础. 与其他的一些任务如 SLAM 和手势识别等相结合，以及如何在复杂场景中，大规模地捕获这些运动变化，是三维物体跟踪的重要研究方向.

三维物体跟踪同样面临诸多挑战，如在物体或相机快速运动产生模糊时，前景的轮廓难以被准确提取；在前背景颜色相似以及传感器噪声存在时，统计分割模型的效果会下降；在前景物体被大面积遮挡时，难以提取足够可靠的轮廓；此外，无纹理对称物体的跟踪仍是一个难点. 上述问题会导致三维物体跟踪算法的精度下降甚至跟踪失败. 三维物体跟踪算法目前仍需要研究者不断改进，以使得其以更高效、鲁棒的方式出现在人们的生活应用中.

参考文献(References):

[1] Lepetit V, Fua P. Monocular model-based 3D tracking of rigid objects[M]. Hanover: Now Publishers Inc, 2005

[2] Marchand E, Uchiyama H, Spindler F. Pose estimation for augmented reality: a hands-on survey[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(12): 2633-2651

[3] Peng S D, Zhou X W, Liu Y, *et al*. PVNet: pixel-wise voting network for 6DoF object pose estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3212-3223

[4] Kirillov Jr A. An introduction to Lie groups and Lie algebras[M]. Cambridge: Cambridge University Press, 2008

[5] Varadarajan V S. Lie groups, Lie algebras, and their representations[M]. New York: Springer, 1984

[6] Murray R M, Li Z X, Sastry S S. A mathematical introduction to robotic manipulation[M]. Boca Raton: CRC Press, 1994

[7] He Xian, Li Jiachen, Jin Li, *et al*. A synthetic dataset and performance evaluation for 3D template tracking[J]. Chinese Journal of Computers, 2022, 45(3): 585-600(in Chinese)
(何弦, 李佳宸, 金立, 等. 三维模板跟踪的基准合成数据集构建及算法评估[J]. 计算机学报, 2022, 45(3): 585-600)

[8] Choi C, Christensen H I. RGB-D object tracking: a particle filter approach on GPU[C] //Proceedings of the IEEE/RSJ Inter-

national Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2013: 1084-1091

[9] Tjaden H, Schwanecke U, Schömer E, *et al*. A region-based Gauss-Newton approach to real-time monocular multiple object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1797-1812

[10] Wu P C, Lee Y Y, Tseng H Y, *et al*. [POSTER] A benchmark dataset for 6DoF object pose tracking[C] //Proceedings of the IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2017: 186-191

[11] Xiang Y, Schmidt T, Narayanan V, *et al*. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes[EB/OL]. [2023-02-15]. https://www.roboticsproceedings.org/rss14/p19.html

[12] Li J C, Wang B, Zhu S Q, *et al*. BCOT: a markerless high-precision 3D object tracking benchmark[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 6687-6696

[13] Shapiro R. Direct linear transformation method for three- dimensional cinematography[J]. Research Quarterly. American Alliance for Health, Physical Education and Recreation, 1978, 49(2): 197-205

[14] Lepetit V, Moreno-Noguer F, Fua P. EP*n*P: an accurate $O(n)$ solution to the P*n*P problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166

[15] Penate-Sanchez A, Andrade-Cetto J, Moreno-Noguer F. Exhaustive linearization for robust camera pose and focal length estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2387-2400

[16] Kato H, Billinghurst M. Marker tracking and HMD calibration for a video-based augmented reality conferencing system[C] //Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality. Los Alamitos: IEEE Computer Society Press, 1999: 85-94

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

[18] Bay H, Ess A, Tuytelaars T, *et al*. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359

[19] Rublee E, Rabaud V, Konolige K, *et al*. ORB: an efficient alternative to SIFT or SURF[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2564-2571

[20] Azarbayejani A, Pentland A P. Recursive estimation of motion, structure, and focal length[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(6): 562-575

[21] Beardsley P A, Zisserman A, Murray D W. Sequential updating of projective and affine structure from motion[J]. International Journal of Computer Vision, 1997, 23(3): 235-259

[22] Simon G, Fitzgibbon A W, Zisserman A. Markerless tracking using planar structures in the scene[C] //Proceedings of the IEEE and ACM International Symposium on Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2000: 120-128

[23] Neumann U, You S. Natural feature tracking for augmented reality[J]. IEEE Transactions on Multimedia, 1999, 1(1): 53-64

[24] Drummond T, Cipolla R. Real-time visual tracking of complex structures[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 932-946

[25] Marchand E, Bouthemy P, Chaumette F, *et al*. Robust real-time visual tracking using a 2D-3D model-based approach[C] //Proceedings of the Seventh International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 1999: 262-268

[26] Vacchetti L, Lepetit V, Fua P. Stable real-time 3D tracking using online and offline information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(10): 1385-1391

[27] Gauglitz S, Höllerer T, Turk M. Evaluation of interest point detectors and feature descriptors for visual tracking[J]. International Journal of Computer Vision, 2011, 94(3): 335-360

[28] Wagner D, Reitmayr G, Mulloni A, *et al*. Pose tracking from natural features on mobile phones[C] //Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2008: 125-134

[29] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C] //Proceedings of the 7th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1981: 674-679

[30] Zhang L, Curless B, Hertzmann A, *et al*. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo[C] //Proceedings of the 9th International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2003: 618-625

[31] Chambolle A, Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging[J]. Journal of Mathematical Imaging and Vision, 2011, 40(1): 120-145

[32] Crivellaro A, Lepetit V. Robust 3D tracking with descriptor fields[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 3414-3421

[33] Seo B K, Wuest H. A direct method for robust model-based 3D object tracking from a monocular RGB image[C] //Proceedings of the European Conference on Computer Vision Workshops. Heidelberg: Springer, 2016: 551-562

[34] Lowe D G. Robust model-based motion tracking through the integration of search and estimation[J]. International Journal of Computer Vision, 1992, 8(2): 113-122

[35] Klein G, Murray D W. Full-3D edge tracking with a particle filter[C] //Proceedings of the British Machine Vision Conference. London: British Machine Vision Association, 2006: 1119-1128

[36] Harris C, Stennett C. RAPID-a video rate object tracker[C] //Proceedings of the British Machine Vision Conference. London: British Machine Vision Association, 1990: 151-156

[37] Wang G F, Wang B, Zhong F, *et al*. Global optimal searching for textureless 3D object tracking[J]. The Visual Computer, 2015, 31(6-8): 979-988

[38] Wang B, Zhong F, Qin X Y. Pose optimization in edge distance field for textureless 3D object tracking[C] //Proceedings of the Computer Graphics International Conference. New York: ACM Press, 2017: Article No.32

[39] Wang B, Zhong F, Qin X Y. Robust edge-based 3D object tracking with direction-based pose validation[J]. Multimedia Tools and Applications, 2019, 78(9): 12307-12331

[40] Huang H, Zhong F, Sun Y Q, *et al*. An occlusion-aware edge-based method for monocular 3D object tracking using edge confidence[J]. Computer Graphics Forum, 2020, 39(7): 399-409

[41] Sun X L, Zhou J X, Zhang W L, *et al*. Robust monocular pose tracking of less-distinct objects based on contour-part model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(11): 4409-4421

[42] Dong Y C, Ji L L, Wang S B, *et al*. Accurate 6DOF pose tracking for texture-less objects[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(5): 1834-1848

[43] Tian X H, Lin X R, Zhong F, *et al*. Large-displacement 3D object tracking with hybrid non-local optimization[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 627-643

[44] Schmaltz C, Rosenhahn B, Brox T, *et al*. Region-based pose tracking[C] //Proceedings of the Third Iberian Conference on Pattern Recognition and Image Analysis. Heidelberg: Springer, 2007: 56-63

[45] Schmaltz C, Rosenhahn B, Brox T, *et al*. Region-based pose tracking with occlusions using 3D models[J]. Machine Vision and Applications, 2012, 23(3): 557-577

[46] Rosenhahn B, Brox T, Weickert J. Three-dimensional shape knowledge for joint image segmentation and pose tracking[J]. International Journal of Computer Vision, 2007, 73(3): 243-262

[47] Kass M, Witkin A, Terzopoulos D. Snakes: active contour models[J]. International Journal of Computer Vision, 1988, 1(4): 321-331

[48] Osher S, Sethian J A. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations[J]. Journal of Computational Physics, 1988, 79(1): 12-49

[49] Prisacariu V A, Reid I D. PWP3D: real-time segmentation and tracking of 3D objects[J]. International Journal of Computer Vision, 2012, 98(3): 335-354

[50] Tjaden H, Schwanecke U, Schömer E. Real-time monocular segmentation and pose tracking of multiple objects[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 423-438

[51] Tjaden H, Schwanecke U, Schömer E. Real-time monocular pose estimation of 3D objects using temporally consistent local color histograms[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 124-132

[52] Zhong L S, Zhao X L, Zhang Y, *et al*. Occlusion-aware region-based 3D pose tracking of objects with temporally consistent polar-based local partitioning[J]. IEEE Transactions on Image Processing, 2020, 29: 5065-5078

[53] Huang H, Zhong F, Qin X Y. Pixel-wise weighted region-based 3D object tracking using contour constraints[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(12): 4319-4331

[54] Stoiber M, Pfanne M, Strobl K H, *et al*. A sparse Gaussian approach to region-based 6DoF object tracking[C] //Proceedings of the 15th Asian Conference on Computer Vision. Heidelberg: Springer, 2021: 666-682

[55] Stoiber M, Pfanne M, Strobl K H, *et al*. SRT3D: a sparse region-based 3D object tracking approach for the real world[J]. International Journal of Computer Vision, 2022, 130(4): 1008-1030

[56] Wang Q F, Zhou J X, Li Z, *et al*. Robust and accurate monocular pose tracking for large pose shift[J]. IEEE Transactions on Industrial Electronics, 2023, 70(8): 8163-8173

[57] Stoiber M, Sundermeyer M, Triebel R. Iterative corresponding geometry: fusing region and depth for highly efficient 3D tracking of textureless objects[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 6845-6855

[58] Zhong L S, Zhang L. A robust monocular 3D object tracking method combining statistical and photometric constraints[J]. International Journal of Computer Vision, 2019, 127(8): 973-992

[59] Liu F L, Wei Z Z, Zhang G J. An off-board vision system for relative attitude measurement of aircraft[J]. IEEE Transactions on Industrial Electronics, 2022, 69(4): 4225-4233

[60] Li J C, Zhong F, Xu S H, *et al*. 3D object tracking with adaptively weighted local bundles[J]. Journal of Computer Science and Technology, 2021: 36(3): 555-571

[61] Li J C, Song X Q, Zhong F, *et al*. Fast 3D texture-less object tracking with geometric contour and local region[J]. Computers & Graphics, 2021, 97: 225-235

[62] Tan D J, Tombari F, Ilic S, *et al*. A versatile learning-based 3D temporal tracker: scalable, robust, online[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 693-701

[63] Tan D J, Navab N, Tombari F. Looking beyond the simple scenarios: combining learners and optimizers in 3D temporal tracking[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(11): 2399-2409

[64] Manhardt F, Kehl W, Navab N, *et al*. Deep model-based 6D pose refinement in RGB[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 833-849

[65] Li Y, Wang G, Ji X Y, *et al*. DeepIM: deep iterative matching for 6D pose estimation[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 695-711

[66] Iwase S, Liu X Y, Khirodkar R, *et al*. RePOSE: fast 6D object pose refinement via deep texture rendering[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 3283-3292

[67] Xu Y, Lin K Y, Zhang G F, *et al*. RNNPose: recurrent 6-DoF object pose refinement with robust correspondence field estimation and pose optimization[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 14860-14870

[68] Wang H, Sridhar S, Huang J W, *et al*. Normalized object coordinate space for category-level 6D object pose and size estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 2637-2646

[69] Tian M, Ang M H, Lee G H. Shape prior deformation for categorical 6D object pose and size estimation[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 530-546

[70] Chen K, Dou Q. SGPA: structure-guided prior adaptation for category-level 6D object pose estimation[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 2753-2762

[71] Lin H B, Liu Z C, Cheang C, *et al*. SAR-Net: shape alignment and recovery network for category-level 6D object pose and size estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 6697-6707

[72] Wang C, Martín-Martin R, Xu D F, *et al*. 6-PACK: category-level 6D pose tracker with anchor-based keypoints[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2020: 10059-10066

[73] Weng Y J, Wang H, Zhou Q, *et al*. CAPTRA: category-level pose tracking for rigid and articulated objects from point clouds[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13189-13198

[74] Wen B W, Bekris K. BundleTrack: 6D pose tracking for novel objects without instance or category-level 3D models[C] //Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2021: 8067-8074

[75] Hinterstoisser S, Holzer S, Cagniart C, *et al*. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 858-865

[76] Hinterstoisser S, Lepetit V, Ilic S, *et al*. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[C] //Proceedings of the 11th Asian Conference on Computer Vision. Heidelberg: Springer, 2012: 548-562

[77] Hinterstoisser S, Cagniart C, Ilic S, *et al*. Gradient response maps for real-time detection of textureless object[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(5): 876-888

[78] Su Y Z, Saleh M, Fetzer T, *et al*. ZebraPose: coarse to fine surface encoding for 6DoF object pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 6728-6738

[79] Xu Z L, Zhang Y C, Chen K, *et al*. BiCo-Net: regress globally, Match locally for robust 6D pose estimation[C] //Proceedings of the 31st International Joint Conference on Artificial Intelligence. Freiburg: International Joint Conferences on Artificial Intelligence, 2022: 1509-1515