计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200095

面向人机协同的物体姿态估计帧间稳定性优化方法

穆逢君[1]　邱　静[1]　陈路锋[2]　黄　瑞[2]　周　林[3]　于功敬[3]
1 电子科技大学机械与电气工程学院　成都 611731
2 电子科技大学自动化工程学院　成都 611731
3 国防科技工业自动化测试创新中心　北京 100041
(mufengjun260@gmail.com)

摘　要　现有的物体姿态估计方法无法提供具有帧间稳定性的估计姿态,导致将其结果直接用于增强现实等可视化场景时会引起画面抖动,不适用于人机协同等应用场景。文中提出了一种包含多种方式的物体姿态估计优化方法,通过对原始姿态估计方法的损失函数的改进,并使用因果滤波的方法优化姿态估计结果,以获得具有稳定性的估计姿态。此外,为完善对姿态估计方法稳定程度的评价体系,文中提出了直接偏差距离 DBD、方向反转率 DRR 与平均位移角 ADA 3 种评价指标,可以从多个角度对物体姿态估计方法的帧间稳定性进行评价。最后,使用 YCB-STB 数据集作为测试样本,并将所提方法与未经优化的原始方法进行对比测试。结果表明,所提方法可在不引入额外资源开销的情况下提高现有物体姿态估计方法的帧间稳定性,且对原始方法的准确率影响较小,满足了人机协同场景对物体姿态估计结果的需求。

关键词:物体姿态估计;人机协同;损失函数;因果滤波
中图法分类号　TP242.6

Optimization Method for Inter-frame Stability of Object Pose Estimation for Human-Machine Collaboration

MU Feng-jun[1],QIU Jing[1],CHEN Lu-feng[2],HUANG Rui[2],ZHOU Lin[3] and YU Gong-jing[3]
1 School of Mechanical and Electrical Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China
2 School of Automation Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China
3 Innovation Center of Automated Testing for Science,Technology and Industry for National Defence,Beijing 100041,China

Abstract　Existing object pose estimation methods cannot provide estimated poses with inter-frame stability. As a result,when the results are directly used in visualization scenarios such as augmented reality,it will cause screen jitter,so it's not suitable enough for application scenarios such as human-machine collaboration. This paper proposes an object pose estimation optimization method that includes multiple methods. By improving the loss function of the original pose estimation method and using causal filtering to optimize the pose estimation result,a stable estimated pose can be obtained. In addition,in order to consummate the evaluation system of the degree of stability of the pose estimation method,this paper proposes three evaluation indicators:the direct deviation distance DBD,the direction reversal rate DRR and the average displacement angle ADA,which can evaluate the object pose estimation method from multiple viewpoints. Finally,the YCB-STB dataset is used to test,and the method is compared with the original method without optimization. The results show that the proposed method can improve the inter-frame stability of the existing object pose estimation methods without introducing additional resources,and has a small impact on the accuracy of the original method,which satisfies the requirement of object attitude estimation in human-machine collaborative scene.

Keywords　Object pose estimate,Human-machine collaboration,Loss function,Causal filtering

1　引言

近年来,随着增强现实技术的快速发展与低成本传感器的普及,人机协同的应用场景也快速扩展,如协同装配、辅助驾驶、机器外肢体等,这些场景通常需要使用增强现实技术将机器感知获取的虚拟环境信息以可视化的形式传达给操作

到稿日期:2020-12-11　返修日期:2021-03-14　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:中央高校基本科研业务费专项资金(ZYGX2019Z010)
This work was supported by the Fundamental Research Funds for the Central Universities(ZYGX2019Z010).
通信作者:邱静(qiujing@uestc.edu.cn)

者，以实现操作者视角下虚拟环境与真实世界的信息融合。

在虚拟环境信息的获取与融合过程中，将实际物体的姿态估计结果使用三维重建的方式展示在操作者视角下[1]。然而，现有的物体姿态估计方法存在估计误差，且姿态估计的误差方向随机，不具有时间上的稳定性，这些因素导致了增强现实画面中可视化的物体姿态信息的不稳定现象。研究结果[2]表明，增强现实中不稳定的画面易引起操作者疲劳、眩晕等不适，从而增加操作者的身体负担，且对使用体验的危害程度显著大于画面帧率、反馈延迟等因素。因此，不稳定的姿态估计结果不利于增强现实等技术的实际应用，阻碍了人机协同应用场景的推广。

经调研，现有物体姿态估计方法的主要提升方向为估计结果的准确性，但人机协同等场景更加注重估计结果的稳定性。因此，本文提出了一种结合损失函数优化方法与因果滤波的稳定性优化方法，利用训练过程中的最优化原理，以特定方向优化方法的网络参数，并使用滤波方法对结果进一步优化。为验证所提方法的性能，本文设计了多种针对物体姿态估计结果稳定性的评价方法，并对使用了不同优化类型的物体姿态估计方法进行了对比。实验结果表明，在基于 YCB-STB 数据集的评估中，使用本文方法可以得到具有稳定性的物体姿态估计结果。

2 相关工作

物体姿态估计方法[3]根据处理方式可分为基于对应关系的方法（Correspondence-based method）、基于模板匹配的方法（Template-based method）以及基于投票的方法（Voting-based method）。根据输入数据类型的不同，每类方法均可分为基于二维图像的方法（2D image-based methods）与基于三维点云的方法（3D point cloud-based methods）。

2.1 基于对应关系的方法

基于对应关系的方法[3]通过处理输入数据（二维图像或三维点云）与给定的物体三维模型之间的特征点对应关系，估计出物体相对于相机的姿态。

基于二维图像的方法适用于纹理丰富的对象，通过生成 2D-3D 匹配特征点对，并利用基于 RANSAC 的 PnP 算法[4]处理匹配特征点的几何关系，即可获得物体坐标系与相机坐标系的相对关系。常用的二维特征点有 FAST 特征点[5]、SURF 特征点[6]与 ORB 特征点[7]。但是，这类方法需要待估计物体表面具有丰富的纹理信息，否则难以提取到二维特征点。因此该类方法不适用于纹理不丰富的简单物体。随着深度神经网络的发展，越来越多的研究表明，利用神经网络可以提取到更具代表性的特征点。SuperPoint[8]提供了一种可以检测特征点，并根据获取的特征点生成特征点描述子的自监督框架，其可同时检测更多特征点并提高特征点匹配的准确性。此外，Hu 等[9]使用深度神经网络替代传统的 PnP 算法，可直接从一组 2D-3D 匹配特征点中回归出物体姿态的估计结果。

基于三维点云的方法通过直接匹配三维几何特征点，消除了纹理不足给特征点带来的影响。三维几何特征点可用于查找物体的部分点云与完整点云之间的对应关系，以估计物体的姿态。但是，该类方法需要物体表面具有丰富的几何特征，否则难以提取到足够的三维特征点。文献[10]提出了一种针对点云的三维几何特征描述符的自监督学习方法，无需标注数据即可得到较好的效果。

2.2 基于模板匹配的方法

基于模板匹配的方法[3]通过将输入与标有物体姿态的模板进行匹配，在模板库中查找出最为相似的模板，并估计物体的姿态。

与基于对应关系的方法类似，基于模板匹配的方法按照输入类型的不同，可分为基于二维图像的方法与基于三维点云的方法。

基于二维图像的方法使用物体的完整 3D 模型在各个方向的投影作为模板，并使用传统方法或基于学习的方法查找出与输入图像最相似的模板图像，然后根据模板图像附加的姿态真值，生成物体估计姿态。Hinterstoisser 等提出的 LineMod[11]是一种扩展梯度方向进行模板匹配并用一组有限的模板描述物体三维信息的方法，该工作提供了 2D 模板匹配方法的通用思路。此外，PoseCNN[12]可以直接从输入图像中估计物体的位置偏移量，并回归四元数来计算物体在三维空间中的旋转，该方法可以被视为隐含在经过训练的神经网络参数中的模板匹配方法。

基于三维图像的方法通过对点云进行旋转与平移，使部分点云与全部点云最优对齐，以得到物体的姿态。Go-ICP[13]使用了全局对齐的方法，可以提供具有较大范围的初始姿态，并对噪声具有鲁棒性。由于全局对齐方法非常耗时，因此 3D 模板匹配方法通常使用局部对齐方法，例如，以迭代最近点算法（Iterative Closest Points，ICP）[14]为代表的传统方法和以 DeepICP[15]为代表的基于学习的方法。G2L-Net[16]可以直接通过从 RGB-D 图像中提取到的粗粒度点云回归得到物体的估计姿态。

2.3 基于投票的方法

基于投票的方法[3]中每个像素与点云中的点均会对物体姿态进行投票，并将最终的投票结果作为估计姿态。投票方法可分为直接投票与间接投票。

直接投票方法意味着每个像素或点云中的点直接对物体的估计姿态进行投票。该类方法可以利用图像中的局部信息约束姿态输出的可能结果，因此可以使用投票的方式估计被部分遮挡物体的姿态。Tejani 等[17]提出了隐类霍夫森林框架，通过将多个局部区域的投票聚类成相互一致的估计姿态来消除检测误差，提升了对物体部分遮挡情况的鲁棒性。DenseFusion[18]使用异构网络得到来自 RGB 数据与点云数据的特征，并进行稠密融合构建像素级特征，避免了直接构建全局特征的方法易受到遮挡影响的缺点。

间接投票方法可被视为基于对应关系的方法的扩充，通过对二维特征点与三维特征点进行投票，得到基于投票的对应关系，并将其用于估计物体的姿态。该类方法通常利用深

度学习强大的特征表示能力来预测更优的投票结果。例如，PVNet[19]与PVN3D[20]使用了像素投票网络(Pixel-wise Voting Network，PVN)，分别对二维特征点与三维特征点进行投票，找到相应的特征点对应关系，并用于估计物体姿态。6-PACK[21]利用了基于学习的方法，用少量的三维特征点来简洁地表示物体，该方法可用于基于RGB-D数据的类别物体的姿态跟踪。

此外，6-PACK[21]与文献[22]提出的方法为姿态跟踪方法，通过估计帧间物体姿态的相对关系，并基于零时刻的初始姿态进行逐帧姿态变换，可以得到任意时刻的物体姿态。由于姿态跟踪方法依赖帧间物体相对姿态的估计结果，因此通常具有较好的帧间稳定性。例如，文献[23]利用线性回归的方法在多个候选结果中进行选择，并使用高斯滤波器消除噪声，可以得到更加平滑的结果。se(3)-TrackNet[24]通过对输入数据与渲染数据的实时比对，仅使用合成数据进行训练即可完成对真实数据中物体的姿态跟踪。PoseRBPF[25]在Rao-Blackwellized粒子滤波框架内实现的物体姿态跟踪方法可以有效估计物体姿态在旋转与平移上的完整分布，为机器人抓取等任务保留了足够的后验信息。但是，姿态跟踪方法在长期的跟踪任务中存在增量累计误差，因此需要频繁对物体姿态进行重初始化，进而导致姿态跟踪方法相对于姿态估计方法在应用于增强现实等场景时存在着画面抖动幅度随时间增大、重初始化时具有较大的帧间姿态估计差异的劣势。

综上所述，现有物体姿态估计方法的主要提升方向为估计结果的准确性，但人机协同等场景对帧间稳定性具有较高的要求。因此，本文针对人机协同等需要稳定的物体姿态估计结果的场景进行研究，其主要贡献如下：

1)提出了一种利用带有DBD(Direct Bias Distance)损失优化项损失函数的优化方法，可基于现有物体姿态估计方法，在不降低原方法运行效率的情况下，提高物体姿态估计结果的稳定性。

2)针对误差向量在各方向上的分量满足高斯分布，且变化频率高于姿态真值的特点，本文提出了基于因果滤波的姿态估计稳定性优化方法，可在不影响准确率的同时减少相邻帧间姿态估计结果不一致的情况发生，且引入的时延不会影响姿态估计结果。

3 方法

本文的目标是对任意基于学习的物体姿态估计方法进行稳定性改进，并对结果进行针对性优化，以提高物体姿态估计方法的帧间稳定性，进而提升其在增强现实等人机协同场景下的应用效果。本文工作基于DenseFusion方法[18](简称DF方法)，通过优化方法模型中的参数和对姿态估计结果进行因果滤波两种优化方法，构建新优化方法。

图1(a)中原始方法模型输出的结果为原始姿态估计结果，该结果未针对稳定性进行优化。使用本文提出的优化损失函数对原始方法进行参数优化，得到的优化方法模型可以输出优化后的姿态估计结果，相较于原始姿态估计结果，该结果具有更好的稳定性。图1(b)中的因果滤波模块可对任意姿态估计结果进行处理，输出比输入更加稳定的滤波后的姿态估计结果。实验部分使用原始姿态估计结果与优化模型的姿态估计结果作为输入进行测试，滤波后的姿态估计结果比输入具有更好的稳定性。

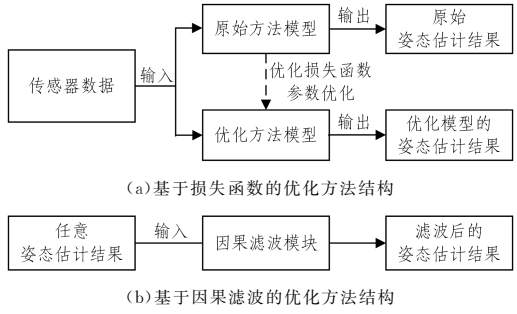


图1 本文提出的两种优化方法与原方法的结构关系

Fig. 1 Relationship between two optimization methods proposed in this paper and original method

物体姿态估计结果通常使用变换矩阵 $\boldsymbol{p}=[\boldsymbol{R}|\boldsymbol{t}]$ 表示[26]，其中旋转矩阵 $\boldsymbol{R}\in SO(3)$ 与位移向量 $\boldsymbol{t}\in R^3$ 分别表示两坐标系之间的旋转关系与位移关系。如图2所示，通过定义变换矩阵 $\boldsymbol{p}$，可以唯一确定物体坐标系与相机坐标系之间的变换关系，即唯一表示物体相对于相机的姿态。

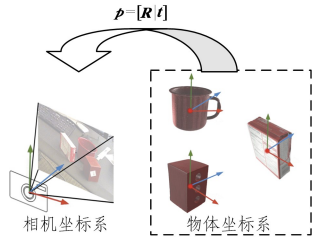


图2 使用变换矩阵确定坐标系之间的变换关系

Fig. 2 Determine transformation relationship with transformation matrix

3.1 DenseFusion介绍

DF使用像素级图像分割的方法获取物体在输入RGB图像与深度图中的掩膜，并通过PSPNet[27]与PointNet[28]分别获取物体的色彩特征与几何特征，用于生成与随机抽样产生的采样像素对应的融合特征。之后，DF使用CNN处理融合特征，回归得到与每个像素对应的四元数 $\boldsymbol{R}$，平移向量 $\boldsymbol{t}$ 和该像素的置信度 C。最后，DF按照置信度最大原则进行投票，选择权重最高的采样像素，将与该像素对应的物体姿态 $\hat{\boldsymbol{p}}=[\hat{\boldsymbol{R}}|\hat{\boldsymbol{t}}]$ 作为物体相对于相机的估计姿态。

3.2 评价方法

对于物体姿态估计任务，通常使用最近平均点云距离(Average Distance，ADD)[8]评价方法的估计误差，其计算方法如下：

$$ADD=\frac{1}{m}\sum_{x\in M}\|(\hat{\boldsymbol{R}}\boldsymbol{x}+\hat{\boldsymbol{t}})-(\boldsymbol{R}\boldsymbol{x}+\boldsymbol{t})\| \tag{1}$$

其中，M 为对应模型空间采样点的集合。ADD表示位于估计姿态与真值姿态下的物体点云间对应点的平均距离。

此外，本文提出直接偏差距离（Direct Bias Distance，DBD）、方向反转率（Direction Reversal Rate，DRR）以及平均偏移角（Average Deviation Angle，ADA）3 个指标，用于评价姿态估计结果不同粒度的稳定性。

定义 1 误差向量 $\boldsymbol{V}$ 表示估计结果 $\hat{\boldsymbol{p}}$ 与真值（Ground Truth，GT）$\boldsymbol{p}$ 的差值。由于非对称物体在坐标系各方向上的分布存在差异，难以仅通过 $\hat{\boldsymbol{p}}$ 与 $\boldsymbol{p}$ 进行误差的准确度量。因此，本文引入随机采样的方法，在物体模型内部及表面取 M 个空间点组成点云 $\{\boldsymbol{x}_h, h=1,\cdots,M\}$ 后，使用变换矩阵将其变换至相机坐标空间中。误差向量 $\boldsymbol{V}$ 为点云间对应点在各方向分量上的平均误差，计算公式如下：

$$\boldsymbol{V}=\frac{1}{M}\sum_{h=1}^{M}[(\hat{\boldsymbol{R}}\boldsymbol{x}_h+\hat{\boldsymbol{t}})-(\boldsymbol{R}\boldsymbol{x}_h+\boldsymbol{t})] \tag{2}$$

其中，$\boldsymbol{x}_h$ 为第 h 个空间采样点。

定义 2 直接偏差距离 DBD 为姿态估计结果中相邻帧的误差向量差的 2-范数，用于度量两误差向量在尺度与方向上的差异水平。其计算公式如下：

$$DBD(\boldsymbol{V}_{\text{pre}},\boldsymbol{V}_{\text{cur}})=\|\boldsymbol{V}_{\text{cur}}-\boldsymbol{V}_{\text{pre}}\| \tag{3}$$

其中，$\boldsymbol{V}_{\text{pre}}$ 与 $\boldsymbol{V}_{\text{cur}}$ 分别为上一帧与当前帧的误差向量。

定义 3 方向反转率 DRR 用于度量姿态估计结果中相邻帧误差向量方向偏差角大于 90°的频率，计算公式如下：

$$DRR=\frac{C_{\text{rev}}}{C_{\text{total}}-1} \tag{4}$$

其中，C_{rev} 为样本中相邻两帧方向向量夹角大于 90°的次数，C_{total} 为总样本帧数。

定义 4 平均偏移角 ADA 为姿态估计结果中相邻帧的误差向量夹角的均值，用于表示视频样本中相邻帧误差向量偏移角度的平均水平。其计算公式如下：

$$ADA=\frac{1}{C_{\text{total}}-1}\sum\langle\boldsymbol{V}_{\text{pre}},\boldsymbol{V}_{\text{cur}}\rangle \tag{5}$$

其中，$\langle\boldsymbol{V}_{\text{pre}},\boldsymbol{V}_{\text{cur}}\rangle$ 表示上一帧与当前帧的两误差向量的夹角。

综上，本文提出了一种相邻帧间姿态估计差异的表示方法与 3 种评价指标，从多个角度评价了物体姿态估计结果的稳定性。ADD 距离度量了估计姿态在空间距离上相对于真值姿态的偏差。误差向量引入了对误差方向的评价，可以同时度量估计姿态与真值姿态下物体点云在距离与方向上的差异。定义 2、定义 3 与定义 4 在定义 1 的基础上引入了时序上的评价方法，使用相邻帧间误差向量的关系作为度量，用于评价相邻帧的姿态估计结果的误差在方向与距离上的稳定性。各度量方式的特点描述如下：

1）误差向量 $\boldsymbol{V}$ 在常用于评价姿态估计准确性的 ADD 的基础上扩充了方向信息，为姿态估计结果在三维空间中的偏差量提供了表示方法。

2）直接偏差距离 DBD 提供了一种简洁的方法，用以度量误差向量在相邻帧中的偏移程度。

3）方向反转率 DRR 以较粗的粒度评价了相邻帧的估计误差在方向上的稳定性，对姿态估计结果在增强现实等场景的应用有着较高的参考价值。

4）平均偏移角 ADA 以更加全局的方式评价了误差向量在完整的样本序列上的稳定性。实际使用时 ADA 可与 DRR 互为补充，更加全面地评价姿态估计方法的时间一致性水平。

本文实验使用上述稳定性度量方法，与现有的物体姿态估计方法进行对比测试，并对各度量指标的结果进行分析。

3.3 基于损失函数的优化方法

经测试[18]，DF 对被遮挡物体同样具有较好的姿态估计精确度。本文实验部分对相邻帧的姿态估计结果 $\boldsymbol{P}_{\text{pre}}$ 与 $\boldsymbol{P}_{\text{cur}}$ 的误差方向进行了对比，DF 有 35.04% 的相邻帧误差方向的夹角大于 90°，即姿态估计结果发生跳变，不利于实际应用于增强现实等需要较高姿态估计稳定性的人机协同场景。

使用将误差向量在帧间的改变纳入损失函数的方法，可在不影响 DF 的姿态估计效果的前提下，避免相邻帧间姿态估计结果不稳定的问题。DF 的原有损失函数为：

$$L_j=\frac{1}{M}\sum_i[(\hat{\boldsymbol{R}}_j\boldsymbol{x}_i+\hat{\boldsymbol{t}}_j)-(\boldsymbol{R}\boldsymbol{x}_i+\boldsymbol{t})]_{\text{L2}} \tag{6}$$

$$L_{\text{ori}}=\frac{1}{N}\sum_j(L_jc_j-w\log c_j) \tag{7}$$

其中，$\{\boldsymbol{x}_i\mid i=1,2,\cdots,M\}$ 为物体模型中随机空间采样点的集合，$\hat{\boldsymbol{p}}=[\hat{\boldsymbol{R}}_j\mid\hat{\boldsymbol{t}}_j]$ 与 c_j 分别为基于第 j 个采样像素（$j=1,2,\cdots,N$）的姿态估计结果与置信度，w 为用于平衡损失函数的超参数。

DF 原有的损失函数 L_{ori} 的优化方向为降低估计姿态与真值姿态下的点云空间对应点间的平均距离，并未考虑相邻帧的误差向量方向不稳定的问题。因此，为降低相邻帧的误差向量的差异，需要在原损失函数的基础上添加帧间稳定项，用于对模型参数进行优化时的稳定性评估。由于方向反转率为事件发生的频率，且各时刻变化量为阶跃函数，无法保证损失函数的可微性，且其与平均偏移角均仅对相邻帧的估计误差在方向上的稳定性进行评估，并未评估估计结果在相邻帧间的偏移程度。因此，本文选择使用误差向量的直接距离作为帧间稳定项添加至 DF 的原始损失函数中，得到可以对帧间稳定性进行优化的损失函数 L_{stable}，该损失函数如下：

$$L_{\text{stable}}=L_{\text{ori}}+\alpha_{\text{stable}}*DBD(\boldsymbol{V}_{\text{pre}},\boldsymbol{V}_{\text{cur}}) \tag{8}$$

其中，α_{stable} 为超参数，用于调节帧间稳定项的权重。由于相邻帧误差向量越平稳，损失向量末端点在三维空间中的距离就越接近，由最终的损失函数可知，姿态估计的帧间稳定性越高，损失函数 L_{stable} 就越低。

算法 1 给出了对相邻帧的估计姿态 $\hat{\boldsymbol{p}}=[\hat{\boldsymbol{R}}\mid\hat{\boldsymbol{t}}]$ 与真值姿态 $\boldsymbol{p}=[\boldsymbol{R}\mid\boldsymbol{t}]$ 进行运算，使损失函数可以用于优化姿态估计结果的帧间稳定性的方法。

算法 1 带有 DBD 优化项的损失函数

输入：上帧与当前帧的估计姿态与真值姿态

$\hat{\mathbf{p}}_{\text{last}}=[\hat{\mathbf{R}}_{\text{pre}}\mid\hat{\mathbf{t}}_{\text{pre}}]$，$\mathbf{p}_{\text{cur}}=[\mathbf{R}_{\text{pre}}\mid\mathbf{t}_{\text{pre}}]$，

$\hat{\mathbf{p}}_{\text{cur}}=[\hat{\mathbf{R}}_{\text{cur}}\mid\hat{\mathbf{t}}_{\text{cur}}]$，$\mathbf{p}_{\text{cur}}=[\mathbf{R}_{\text{cur}}\mid\mathbf{t}_{\text{cur}}]$；

物体模型采样得到的点云 **PC**

输出：可微的损失函数值 L_{stable}

1. function stableLoss()

2. $\mathbf{V}_{\text{last}}=\text{mean}((\mathbf{PC}*\hat{\mathbf{R}}_{\text{pre}}+\hat{\mathbf{t}}_{\text{pre}})-(\mathbf{PC}*\mathbf{R}_{\text{pre}}+\mathbf{t}_{\text{pre}}))$

3. $\mathbf{V}_{cur}=\text{mean}((\mathbf{PC}*\hat{\mathbf{R}}_{cur}+\hat{\mathbf{t}}_{cur})-(\text{PC}*\mathbf{R}_{cur}+\mathbf{t}_{cur}))$

4. $L_j=\frac{1}{M}\sum_i[(\hat{\mathbf{R}}_j\mathbf{x}_i+\hat{\mathbf{t}}_j)-(\mathbf{R}\mathbf{x}_i+\mathbf{t})]_{L2}$

5. $L_{ori}=\frac{1}{N}\sum_j(L_j c_j-\text{wlog}\ c_j)$

6. $L_{stable}=L_{ori}+\alpha_{stable}*\text{DBD}(\mathbf{V}_{pre},\mathbf{V}_{cur})$

算法 1 以公式序列的形式给出了从原始网络的估计姿态与真值姿态计算带有 DBD 优化项的损失函数的过程。该损失函数算法中各运算过程均可微，可直接与基于学习的原始方法结合，使用 Adam[29] 等优化器对网络参数进行优化。

3.4 基于因果滤波的优化方法

在物体的姿态估计任务中，姿态真值的帧间变化频率较低，估计误差向量的帧间变化频率较高且在各方向上均服从高斯分布，因此各方向上的误差分量可被视为高频的高斯噪声。为了得到更加适用于增强现实等人机协同场景的物体姿态估计方法，需要在保留更多真值信息的同时，对在时序序列中估计姿态的高频噪声进行滤除。基于因果滤波的优化借鉴了数字图像处理中高斯滤波的思想，使用因果系统中的滑动平均滤波器对姿态估计结果进行补偿，提升了姿态估计结果的稳定性。

原始姿态估计方法在时刻 T 的输出为姿态估计结果 $\hat{\boldsymbol{p}}_T=[\hat{\boldsymbol{R}}_{out}|\hat{\boldsymbol{t}}_{out}]$。由于滑动平均滤波器会导致输出滞后，为避免过长的滞后降低姿态估计的应用效果，本文方法将滤波窗口设定为 5，滑动平均滤波器的输出与各时刻估计姿态的关系如下：

$$\boldsymbol{R}_{out}=\text{QuadAvg}(\{\text{Quad}(\hat{\boldsymbol{R}}_{T-i})|i=0,\cdots,4\}) \tag{9}$$

$$\boldsymbol{t}_{out}=\frac{1}{5}\sum_{i=0}^{4}(\hat{\boldsymbol{t}}_{T-i}) \tag{10}$$

$$\boldsymbol{p}_{out}=[\boldsymbol{R}_{out}|\boldsymbol{t}_{out}] \tag{11}$$

其中，$\{\hat{\boldsymbol{R}}_{T-4},\cdots,\hat{\boldsymbol{R}}_T\}$ 与 $\{\hat{\boldsymbol{t}}_{T-4},\cdots,\hat{\boldsymbol{t}}_T\}$ 分别为从 -4 时刻至当前时刻 T 的旋转矩阵序列与位移向量序列；$\boldsymbol{p}_{out}$ 为该方法输出的姿态估计结果；函数 Quad 将旋转矩阵转化为四元数；函数 QuadAvg[30] 给出了多个四元数求均值的方法，其中各四元数权重相同。滑动平滑滤波器的输出相对于输入有 $\frac{N-1}{2}$ 个采样的延迟，因此应用该滤波器对估计姿态进行平滑后，会引入 $\frac{5-1}{2}=2$ 帧的延迟。人机协同场景下的输入帧率为 30 fps，2 帧延迟对应的时间为 0.067 s，因此，该优化方法仅在原有姿态估计方法的基础上增加较小的延迟即可提高估计姿态的稳定性。该方法仅需原始物体姿态估计方法的估计姿态，同时适用于对基于学习或传统方法的方法的优化。

4 实验设置与测试数据

4.1 实验环境

实验在单台服务器上进行，实验硬件环境为 CPU Intel (R) Xeon(R) CPU E5-2620 v3 @ 2.40 GHz，GPU NVIDIA GeForce RTX 2080Ti，内存 128 GB。实验用的软件环境中，操作系统为 Ubuntu 16.04 LTS，Python 版本为 3.6.10，PyTorch 版本为 1.1.0。

4.2 测试数据集 YCB-STB

本文面向增强现实等需要时间一致性的场景，实际使用时输入数据为包含 RGB 图像与深度信息的视频序列。因此，选用 YCB-Video 数据集[8,31] 进行分析，该数据集涵盖了 21 个不规则物体模型(见图 3)，包含 13 万帧采集自真实场景与 8 万帧合成渲染的 RGB-D 数据，其中真实场景数据由 92 个 RGB-D 视频序列组成，且均存在物体遮挡等增大物体姿态估计难度的外部影响因素。

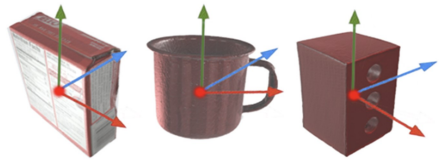


图 3 YCB-Video 数据集提供的部分物体模型

Fig. 3 Several object models from YCB-Video dataset

原始的 YCB-Video 数据集按照随机采样的原则无序抽取了真实数据与合成数据中的部分帧，并划分为训练集、测试集与验证集。由于本文方法的研究重点为人机协同场景下姿态估计方法的帧间稳定性优化方法，因此不能直接使用 YCB-Video 中的划分方式进行训练与评价。由于 YCB-Video 数据集中视频的采样率过高，不利于模型的快速训练与优化，我们选用固定间隔采样的方法，每 6 帧抽取 1 帧，组成可用于姿态估计稳定性评估的完整数据集 YCB-STB，并按照 4∶3∶3 的比例进行训练集、测试集与验证集的划分。YCB-STB 数据集包含 22135 帧采集自真实场景的有序 RGB-D 数据，可用于各类基于时序信息的物体姿态估计方法的训练与测试。

5 结果与讨论

5.1 基于损失函数的优化方法评价

本节分别使用原有损失函数 L_{ori} 与带有 DBD 损失优化项的损失函数 L_{stable}，基于 YCB-STB 数据集进行模型训练，表 1 和表 2 列出了训练得到的原始 DF 方法与优化方法的模型在多种指标下对 YCB-STB 测试集的测试结果。

表 1 原始 DF 方法的测试结果

Table 1 Test results of the original DF method

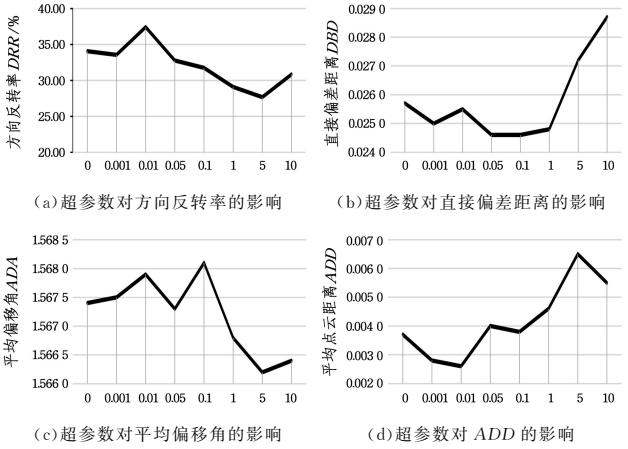
DBD	*DRR*/%	*ADA*	*ADD*
0.0257	35.04	1.5674	0.0037

表 2 使用不同超参数的优化方法的测试结果

Table 2 Test results of optimized model trained with different hyperparameters

超参数 α_{stable}	*DBD*	*DRR*/%	*ADA*	*ADD*
0.001	0.0250	33.55	1.5675	0.0028
0.01	0.0255	37.40	1.5679	0.0026
0.05	0.0246	32.75	1.5673	0.0040
0.1	0.0246	31.75	1.5681	0.0038
1	0.0248	29.10	1.5668	0.0046
5	0.0272	27.66	1.5662	0.0065
10	0.0287	30.83	1.5664	0.0055

表 1 和表 2 使用了 4 种不同的评价指标，对原始 DF 方法与优化方法在 YCB-STB 数据集上的稳定性与精度进行对比测试。由表 1 可见，有 35.04% 的帧与相邻帧的误差向量方向偏差角大于 90°，具有较大的方向反转率，因此，实际使用原始 DF 方法时易引起画面抖动等现象。表 2 中超参数设定为 1 的测试结果中，优化方法的直接偏差距离 DBD、方向反转率 DRR、平均偏转角 ADA 3 项指标分别为 0.024 8，29.10% 与 1.566 8，均优于表 1 中原始 DF 方法的结果，且用于评价准确率的 ADD 为 0.0046，接近表 1 中使用原始 DF 方法测试得到的 ADD。图 4 给出了随着超参数的逐渐增大，各项评价指标的变化趋势。虽然 α_{stable} 为 5 时方向反转率 DRR 与平均偏移角 ADA 均具有测试范围内最优的结果，但其在直接偏差距离 DBD 与 ADD 两项评价指标上结果较差，因此后续实验中选用更为均衡的超参数进行训练，即 α_{stable} 为 1。



(a)超参数对方向反转率的影响 (b)超参数对直接偏差距离的影响

(c)超参数对平均偏移角的影响 (d)超参数对 *ADD* 的影响

图 4 超参数 α_{stable} 对各评价因素的影响

Fig. 4 Influence of hyperparameter α_{stable} on various evaluation factors

该优化方法由于仅对网络训练过程中的损失函数进行调整，并不参与正向推理的过程，因此应用在实际系统时无需额外的计算开销，适用于各类应用场景。在增强现实等需要人机协同的实际应用场景下，操作者是人机协同系统闭环中的重要环节。为保证操作者的使用体验，减少画面跳变等现象的发生，帧间稳定性的优先级高于估计准确性。因此，本优化方法虽然会略微降低估计姿态的准确程度，但其具有更优的估计稳定性，优化后的方法相较于原始方法更加适用于人机协同等实际场景。

综上，本文方法在保证了较好的估计准确性的同时，避免了姿态估计结果的相邻帧间跳变现象的频繁发生，提高了物体姿态估计方法在实际系统中的可用性，可以满足人机协同场景下对物体进行跟踪与操作的姿态估计需求。

5.2 基于因果滤波的优化方法评价

为更好地体现本文提出的两种优化方法的使用效果，分别使用超参数 α_{stable} 为 0(即无优化)与 1.0 时的损失函数 L_{stable} 训练的 DF 网络的输出结果作为滤波器的输入数据，并使用上述提出的多种指标对滤波后的姿态估计结果进行评估。实验结果如表 3 所列。

表 3 基于因果滤波的优化方法测试结果

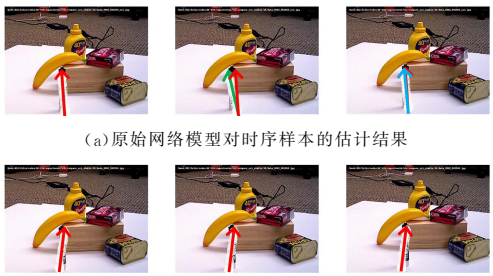
Table 3 Test results of optimization method based on causal filtering

优化方式	*DBD*	*DRR*/%	*ADA*	*ADD*
DF 方法	0.025 7	35.04	1.567 4	0.003 7
损失优化	0.024 8	29.10	1.566 8	0.004 6
因果滤波	0.028 3	31.19	1.566 5	0.003 7
同时优化	0.028 4	27.93	1.566 2	0.004 9

表 3 体现了使用本文提出的两种优化方法对直接偏差距离 DBD、方向反转率 DRR、平均偏移角 ADA 与 ADD 的影响。经过因果滤波，原始 DF 方法的 DRR 从 35.04% 降至 31.19%，且 ADD 未发生变化。因此，基于因果滤波的优化方法可在保持原始物体姿态估计方法准确率的同时，显著降低高斯误差对结果的影响，提升估计姿态的帧间稳定性。此外，同时使用本文提出的两种优化方法的测试表明，基于损失函数的优化方法与基于因果滤波的优化方法耦合度低，未发生互相干扰的情况，可以在实际系统中同时使用。

5.3 讨论

本文在对 4 种评价指标进行对比的同时，结合姿态估计结果与物体模型，按照估计结果旋转平移模型点云后将其投影到输入的 RGB 数据中，生成了物体姿态估计结果在增强现实中的画面效果。图 5(a)与图 5(b)对比了使用原始方法与优化方法对测试集中物体进行姿态估计时，连续 3 帧的估计结果在画面中的显示效果。



(a)原始网络模型对时序样本的估计结果

(b)优化网络模型对时序样本的估计结果

图 5 优化前后物体姿态估计结果的显示效果对比(电子版为彩色)

Fig. 5 Comparison of display effect of object pose estimation results before and after optimization

图 5 各子图中红色箭头方向均一致，图 5(a)中第 1 帧与第 3 帧姿态估计结果基本一致，第 2 帧的物体实际位置并未发生变化，而姿态估计结果与前后两帧在朝向上具有明显差异，因此该估计结果的帧间稳定性较差，难以直接应用在要求稳定性的增强现实等场景中。图 5(b)中 3 帧的估计结果的方向基本相同，不易在可视化时发生抖动等现象，适用于增强现实等场景。因此，使用带有 DBD 损失优化项的损失函数 L_{stable} 训练 DF 网络，得到的优化方法具有帧间稳定性，其物体姿态估计结果不会引起增强现实中画面的抖动，适用于人机协同等场景。

结束语 本文针对现有物体姿态估计方法无法提供人机协同等场景需要的稳定姿态估计结果的问题，设计并实现了对现有物体姿态估计方法进行稳定性优化的方法，包括基于损失函数与基于因果滤波两种方法。基于损失函数的优化方法适用于基于学习的方法，由于改进方式为在原有损失函数

上添加 DBD 优化项，因此该方法适用于各类网络结构。此外，基于因果滤波的优化方法还适用于传统方法，且具有更为广泛的通用性。评估结果表明。本文提出的稳定性优化方法在不引入额外资源开销的情况下能够提高原始物体姿态估计方法的帧间稳定性，且面向实际使用时对姿态估计准确性的不同需求，本文提出的两种方法可以组合使用，以实现姿态估计方法在不同水平的准确率与稳定性间切换，可以满足人机协同等需要较高帧间稳定性的场景对物体姿态估计的实际需求。

未来将结合增强现实技术，研究物体姿态估计方法在人机协同中的应用，并发掘该技术在工程中的实用价值。此外，还需进一步研究同时具有高准确性、高稳定性与零时延等特性的物体姿态估计方法。

参 考 文 献

[1] VAN KREVELEN D W F, POELMAN R. A survey of augmented reality technologies, applications and limitations[J]. International Journal of Virtual Reality, 2010, 9(2): 1-20.

[2] LOUIS T, TROCCAZ J, ROCHET-CAPELLAN A, et al. Is it real? measuring the effect of resolution, latency, frame rate and jitter on the presence of virtual entities[C]// International Conference on Interactive Surfaces and Spaces (ISS). ACM, 2019: 5-16.

[3] DU G, WANG K, LIAN S, et al. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review[J]. Artificial Intelligence Review, 2020, 54(3): 1677-1734.

[4] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: An Accurate O(n) Solution to the PnP Problem[J]. International Journal of Computer Vision (IJCV), 2009, 81(2): 155-166.

[5] ROSTEN E, DRUMMOND T. Fusing Points and Lines for High Performance Tracking[C]// International Conference on Computer Vision (ICCV). IEEE, 2005: 1508-1515.

[6] BAY H, TUYTELAARS T, VAN GOOL L. SURF: Speeded up Robust Features[C]// European Conference on Computer Vision (ECCV). Springer, 2006: 404-417.

[7] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF[C]// International Conference on Computer Vision (ICCV). IEEE, 2011: 2564-2571.

[8] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: Self-Supervised Interest Point Detection and Description[C]// Conference on ComputerVision and Pattern Recognition Workshop (CVPRW). IEEE/CVF, 2018: 224-236.

[9] HU Y, HUGONOT J, FUA P, et al. Segmentation-Driven 6D Object Pose Estimation[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, 2019: 3385-3394.

[10] YUAN Y, HOU J, NÜCHTER A, et al. Self-supervised Point Set Local Descriptors for Point Cloud Registration[J]. arXiv: 2003.05199, 2020.

[11] HINTERSTOISSER S, HOLZER S, CAGNIART C, et al. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes[C]// International Conference on Computer Vision (ICCV). IEEE, 2011: 858-865.

[12] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes[J]. arXiv: 1711.00199, 2017.

[13] YANG J, LI H, CAMPBELL D, et al. Go-ICP: A globally optimal solution to 3D ICP point-set registration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2016, 38(11): 2241-2254.

[14] BESL P J, MCKAY N D. A Method for Registration of 3D Shapes[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). IEEE, 1992: 586-606.

[15] LU W, WAN G, ZHOU Y, et al. DeepICP: An end-to-end deep neural network for 3D point cloud registration[J]. arXiv: 1905.04153, 2019.

[16] CHEN W, JIA X, CHANG H J, et al. G2L-Net: Global to Local Network for Real-time 6D Pose Estimation with Embedding Vector Features[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, 2020: 4233-4242.

[17] TEJANI A, TANG D, KOUSKOURIDAS R, et al. Latent-Class Hough Forests for 3D Object Detection and Pose Estimation [C]// European Conference on Computer Vision (ECCV). Springer, 2014: 462-477.

[18] WANG C, XU D, ZHU Y, et al. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, 2019: 3343-3352.

[19] PENG S, LIU Y, HUANG Q, et al. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, 2019: 4561-4570.

[20] HE Y, SUN W, HUANG H, et al. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, 2020: 11632-11641.

[21] WANG C, MARTÍN-MARTÍN R, XU D, et al. 6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints[C]// International Conference on Robotics and Automation (ICRA). IEEE, 2020: 10059-10066.

[22] DENG X, XIANG Y, MOUSAVIAN A, et al. Self-supervised 6d object pose estimation for robot manipulation[C]// International Conference on Robotics and Automation (ICRA). IEEE, 2020: 3665-3671.

[23] CRIVELLARO A, RAD M, VERDIE Y, et al. Robust 3D Object Tracking from Monocular Images Using Stable Parts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2017, 40(6): 1465-1479.

[24] WEN B, MITASH C, REN B, et al. se(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains[J]. arXiv: 2007.13866, 2020.

[25] DENG X, MOUSAVIAN A, XIANG Y, et al. PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Tracking [J]. arXiv: 1905.09304, 2019.

[26] GAO X, ZHANG T, YAN Q R, et al. 14 Lectures on Visual SLAM: From Theory to Practice[M]. Publishing House of Electronics Industry, 2017.

[27] ZHAO H, SHI J, QI X, et al. Pyramid Scene Parsing Network

[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF, 2017: 2881-2890.

[28] QI C R, SU H, MO K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]// Conference on Computer Vision and Pattern Recognition(CVPR). IEEE/CVF, 2017: 652-660.

[29] KINGMA D P, BA J. ADAM: Adam: A Method for Stochastic Optimization[J]. arXiv: 1412. 6980, 2014.

[30] MARKLEY F L, CHENG Y, CRASSIDIS J L, et al. Averaging quaternions[J]. Journal of Guidance, Control, and Dynamics, 2007, 30(4): 1193-1197.

[31] CALLI B, SINGH A, WALSMAN A, et al. The YCB object and Model set: Towards common benchmarks for manipulation research[C]// International Conference on Advanced Robotics (ICAR). IEEE, 2015: 510-517.

MU Feng-jun, born in 1997, postgraduate. His main research interests include computer vision and human-machine collaboration.

QIU Jing, born in 1977, Ph.D, associate professor. Her main research interests include exoskeleton robot and human factors engineering.