論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Study on Recovering Camera Motion and 3-D Structures from Sequential Monocular Images
著者(和文)	JIANGZijie
Author(English)	Zijie Jiang
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12808号, 授与年月日:2024年6月30日, 学位の種別:課程博士, 審査員:奥富 正敏,塚越 秀行,中臺 一博,田中 正行,原 精一郎,川上 玲
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12808号, Conferred date:2024/6/30, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	システム制御 系 システム制御 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（ 工学 ）
学生氏名： Student's Name	JIANG ZIJIE	審査員主査： Chief Examiner	奥富正敏	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Recovering camera motion and 3D structures from sequential monocular images is a fundamental problem in computer vision that aims to construct a minimal 3D perception system using only a monocular camera. This approach offers advantages in affordability, simplicity, and versatile applicability compared to other 3D perception systems such as Lidar, time-of-flight sensors, and structured light cameras. However, the inherent limitations of monocular cameras in directly measuring 3D information pose significant challenges. The research presented in this thesis addresses these challenges through three primary approaches: traditional geometry-based methods, learning-based data-driven approaches, and methods based on neural implicit representation.

Traditional geometry-based approaches rely on the principles of multi-view geometry to recover 3D structure and camera motion. These methods, such as Structure from Motion (SfM), use feature correspondences between multiple images to estimate camera poses and reconstruct 3D scenes. While effective in many scenarios, these approaches can fail in challenging environments with texture-less scenes, motion blur, or insufficient overlapping views. Traditional SfM and visual simultaneous localization and mapping (vSLAM) have matured over the years, offering robust 2D-2D correspondences and optimized camera poses and 3D structures through bundle adjustment. However, they struggle with non-Lambertian and texture-less scenes, and have limited tolerance for noise in the estimated 2D-2D correspondences.

Learning-based methods leverage large-scale datasets to train models for depth and motion estimation. These approaches have shown strong robustness and adaptability, often outperforming traditional methods in various challenging environments. However, their performance heavily depends on the quality and scale of the training data, and they may struggle with generalization to unseen scenarios. Notably, self-supervised learning has been explored to reduce the reliance on labeled data, utilizing proxy losses such as photometric loss based on differentiable image synthesis. Despite significant advancements, self-supervised methods still face challenges in long-term pose estimation and handling dynamic objects in real-world datasets.

Neural implicit representation approaches provide a more compact and expressive scene representation. These methods, such as neural radiance fields (NeRF), have shown impressive results in expressing both indoor and outdoor scenes. NeRF models a 3D scene as a field function that predicts density and color from spatial coordinates, using differentiable volume rendering to produce rendered images from arbitrary viewpoints. This approach offers a compact and memory-efficient way to represent 3D scenes, yet its application in more specialized domains, such as the medical field, remains under exploration. Moreover, while neural implicit methods have bridged the gap between image collections and 3D reconstructions, they often rely on preprocessed camera poses from traditional SfM techniques.

The thesis addresses several challenges across these approaches and proposes effective methods to enhance robustness, efficiency, and accuracy in 3D reconstruction and motion estimation. First, a robust and efficient SfM pipeline for accurate 3D reconstruction under challenging environments is presented, leveraging camera pose information from visual-inertial odometry (VIO). This method includes a geometric verification process that filters out mismatches by considering the prior geometric configuration of candidate image pairs. By combining VIO and SfM, this approach improves robustness in scenarios with degraded visual information. Second, the thesis introduces a framework called Multi-Layer Fusion Visual Odometry (MLF-VO) for self-supervised learning of depth and ego-motion estimation. This method leverages RGB and inferred depth information in a multi-layer fusion manner, improving the performance of ego-motion estimation in various scenarios. Detailed studies on fusion strategies and design choices demonstrate the framework's advantages.

To leverage real-world, unlabeled datasets with dynamic objects, the thesis integrates 3D scene

flow estimation into the self-supervised framework for camera motion and depth estimation. The proposed model, EMR-MSF, imposes geometric constraints with an ego-motion aggregation module and introduces a motion consistency loss along with a mask regularization loss. These strategies enhance the stability and accuracy of ego-motion estimation. Finally, the thesis applies NeRF to the medical domain, proposing GastroNeRF for rendering photo-realistic images from novel viewpoints within the patient's stomach using monocular gastroscopic data. This method incorporates geometry-based supervision from reconstructed point clouds, addressing performance degradation due to view sparsity in monocular gastroscopy. It improves the quality of novel view synthesis, providing better assistance for clinical diagnosis and intervention.

Extensive experiments demonstrate the effectiveness of the proposed methods. The VIO-aided SfM approach shows significant improvements in 3D reconstruction accuracy in challenging environments compared to traditional SfM methods. The MLF-VO framework outperforms state-of-the-art methods in ego-motion estimation, and the EMR-MSF model achieves superior performance in scene flow estimation and generalization ability. GastroNeRF successfully synthesizes high-quality novel views in gastroscopic applications, validating its potential for practical medical use.

In conclusion, the thesis presents a comprehensive study on recovering camera motion and 3D structures from monocular image sequences, addressing critical challenges in traditional, learning-based, and neural implicit representation approaches. The proposed methods demonstrate significant advancements in robustness, efficiency, and accuracy, with promising applications in robotics, autonomous navigation, and medical imaging. Future work includes further refinement of these methods and exploration of their applications in other specialized domains.