Literature Survey: Deep Reinforcement Learning for Optimal Camera View Selection in 3D Sports Reconstruction

Project Summary

My project addresses the computational challenge in multi-view 3D sports reconstruction by developing a reinforcement learning agent that dynamically selects an optimal subset of camera views. This approach aims to maximize reconstruction quality while minimizing computational costs.

Problem Formulation

The camera view selection problem is inherently sequential, with each selection affecting subsequent choices and the final reconstruction quality. The state space comprises current reconstruction quality metrics, features extracted from available camera views, athletes' positions and movements, occlusion information, and previously selected views. The action space is discrete, representing the selection of one of N available cameras.

The transition function maps the current state and chosen camera view to the next state by updating the reconstruction with the selected view and recalculating quality metrics, features, and occlusion information. The reward function balances reconstruction quality improvements (measured by reduction in reprojection error, increased point cloud density, and reduced occlusions) against computational costs, with a terminal reward based on final reconstruction quality compared to using all views.

Implementation Plan

The project will implement Deep Q-Network (DQN) as the primary RL algorithm, with comparisons against baselines including using all available camera views and random selection. The implementation spans setting up a reconstruction pipeline using COLMAP, implementing the RL environment, training the agent, and evaluating against baselines.

Literature Review

RL-Based View Selection

"Dynamic camera configuration learning for high-confidence active object detection" (Arsénio et al., 2021) Arsénio et al. propose a reinforcement learning approach for optimal camera configuration in active object detection scenarios. They develop a method that dynamically selects and positions cameras to maximize detection confidence while minimizing the number of cameras used. Their system employs a DQN agent that learns to select camera configurations based on scene geometry and object visibility. The state representation includes detection confidence maps and geometric features, while their reward function balances detection performance against resource usage. Our project shares the goal of intelligent camera selection but focuses specifically on 3D reconstruction rather than object detection. We can adapt their approach of balancing performance metrics against resource constraints in our reward function design. Their

work provides valuable insights into how to formulate camera selection as a sequential decision-making problem, though we extend this concept to the domain of dynamic sports scenarios where occlusions and movements create additional challenges. Their results demonstrating significant resource savings while maintaining high detection performance align with our objective of computational efficiency in reconstruction.

"View planning in robot active vision: A survey of systems, algorithms, and applications" (Lu et al., 2021) This comprehensive survey examines next-best-view (NBV) planning techniques, including reinforcement learning approaches. The authors categorize NBV methods based on their objective functions, which typically include information gain, reconstruction accuracy, and coverage. While not specifically using RL, many of the surveyed approaches employ information-theoretic metrics to evaluate potential views. Our project aligns with the NBV paradigm but extends it to the domain of multi-view sports reconstruction. The survey's discussion of information gain metrics could inform our state representation and reward function design. Unlike many NBV approaches that focus on exploration of unknown environments, our project emphasizes efficient reconstruction of dynamic scenes with known camera positions.

"Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks" (Jayaraman & Grauman, 2018) Jayaraman and Grauman present an approach for learning exploratory looking policies that are task-independent. They train a reinforcement learning agent to make view selection decisions that maximize information gain across various tasks. Their approach uses a curiosity-driven reward function that encourages exploration of informative viewpoints. While our project shares the view selection objective, we differ in that our task is specifically defined (3D reconstruction) rather than task-agnostic. However, their information gain formulation could be adapted for our reward function to encourage the selection of views that provide complementary information to already selected views, potentially improving reconstruction efficiency.

"Supervised learning of the next-best-view for 3D object reconstruction" (Mendoza et al., 2020) Mendoza et al. introduce a supervised deep learning approach to the next-best-view problem for 3D object reconstruction. Unlike traditional search-based methods, they employ a three-dimensional convolutional neural network (3D-CNN) that directly predicts the optimal sensor pose from the current state of the reconstruction. Their method includes an innovative algorithm for automatic generation of training datasets and demonstrates faster performance with comparable coverage to state-of-the-art methods. Our project diverges from their approach by using reinforcement learning rather than supervised learning. While they train a 3D-CNN to directly map the current reconstruction state to the best next view, our RL formulation learns a policy through interaction with the environment. Their approach requires generating comprehensive training data covering possible reconstruction scenarios, whereas our RL agent learns from experience. However, their method for encoding the current reconstruction state as input to the neural network could inform our state representation design. Additionally, their approach to automatically generating training data could be adapted to create initial environments for our RL agent to explore during early training phases. Their results demonstrating faster view selection with high coverage are particularly relevant to our goal of balancing reconstruction quality with computational efficiency. The comparative evaluation methodology they employ could also guide our baseline comparison approach. Our project could potentially benchmark against their supervised learning method as an additional baseline.

3D Reconstruction Techniques

"COLMAP: General-Purpose Structure-from-Motion and Multi-View Stereo" (Schönberger & Frahm, 2016) COLMAP is a widely used open-source software for structure-from-motion and multi-view stereo reconstruction. Schönberger and Frahm describe the pipeline's incremental reconstruction approach, feature extraction, matching, and bundle adjustment techniques. The tool provides comprehensive metrics for reconstruction quality assessment. Our project will leverage COLMAP as the underlying reconstruction system, but we aim to enhance it with intelligent view selection. Understanding COLMAP's internal metrics for reconstruction quality is essential for designing our reward function. Our contribution lies in the integration of RL-based view selection with COLMAP's reconstruction capabilities to optimize computational efficiency.

"Multi-View Stereo: A Tutorial" (Furukawa & Hernández, 2015) This tutorial provides a comprehensive overview of multi-view stereo (MVS) techniques, including point cloud, volumetric, and mesh-based approaches. The authors discuss the fundamental challenges in MVS, such as establishing correspondences across views, handling occlusions, and ensuring geometric consistency. Our project directly addresses these challenges through intelligent view selection. Understanding the technical aspects of MVS helps us design appropriate features for our state representation and metrics for our reward function. Unlike traditional MVS approaches that use all available views, our method will adaptively select views based on their contribution to reconstruction quality.

Computational Efficiency in 3D Vision

"Efficient Multi-View Performance Capture of Fine-Scale Surface Detail" (Collet et al., 2015) Collet et al. present a system for efficient multi-view performance capture that balances reconstruction quality with computational constraints. They describe techniques for adaptive mesh refinement and texture mapping that focus computational resources on regions of interest. Our project shares the goal of computational efficiency but approaches it through camera selection rather than adaptive processing. Their discussion of quality-computation tradeoffs provides valuable insights for our reward function design. We could potentially combine our view selection approach with their adaptive processing techniques for even greater efficiency.

"Real-time 3D Reconstruction at Scale using Voxel Hashing" (Nießner et al., 2013) This paper introduces an efficient data structure for large-scale 3D reconstruction that minimizes memory usage and computational overhead. Nießner et al. demonstrate real-time performance for room-scale scenes using a voxel hashing approach. While our project focuses on view selection rather than data structures, understanding efficient reconstruction algorithms informs our computational cost modeling in the reward function. Their real-time processing approach aligns with our stretch goal of extending to streaming scenarios. We could potentially integrate their voxel hashing technique with our view selection method for scalable reconstruction. Applications of RL in Computer Vision

"Reinforcement Learning for Visual Object Detection" (Pirinen & Sminchisescu, 2018) Pirinen and Sminchisescu apply reinforcement learning to object detection by formulating it as a sequential decision problem. Their agent learns to focus computational resources on promising image regions, iteratively refining detection results. Although focused on 2D detection rather than 3D reconstruction, their approach demonstrates the effectiveness of using RL to optimize computational resource allocation in vision tasks. Their state representation design and their progressive refinement strategy could inspire our approach to sequential view selection.

Conclusion and Relation to Our Project

Our proposed RL-based camera view selection for 3D sports reconstruction builds upon and extends several lines of research: RL-based view selection, supervised learning for next-best-view planning, 3D reconstruction techniques, sports-specific vision, and computational efficiency in 3D vision.

Unlike previous work that has typically focused on either view selection for exploration or computational efficiency in reconstruction, our approach uniquely combines these objectives for the specific domain of sports reconstruction. The sequential nature of our formulation allows the RL agent to learn complementary view selection strategies that progressively improve reconstruction quality while managing computational costs.

The distinctive aspects of our project include:

- Domain specificity to sports scenarios with dynamic movement and occlusions
- Explicit modeling of computational costs in the reward function
- Integration with an established reconstruction pipeline (COLMAP)
- Focus on sequential view selection rather than one-time camera placement
- Learning through interaction rather than supervised training on pre-generated data (contrasting with Mendoza et al.)

References

Arsénio, A., Santos-Victor, J., & Bernardino, A. (2021). Dynamic camera configuration learning for high-confidence active object detection. Neurocomputing, 466, 305-316. https://doi.org/10.1016/j.neucom.2021.09.037

Collet, A., Zollhöfer, M., Kim, H.-S., Saragih, J., Prince, S. J. D., & Nießner, M. (2015). High-Quality Capture of Fine-Scale Dynamic Surface Detail. ACM Transactions on Graphics (TOG), 34(6), Article 229, 1–13. https://doi.org/10.1145/2816795.2818066

Furukawa, Y., & Hernández, C. (2015). Multi-View Stereo: A Tutorial. Foundations and Trends® in Computer Graphics and Vision, 9(1–2), 1-148. http://dx.doi.org/10.1561/0600000052

Jayaraman, D., & Grauman, K. (2018). Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8356-8365). https://doi.org/10.1109/CVPR.2018.00872

Lu, F., Xue, F., Wu, S., Zhang, X., & Tan, G. (2021). View planning in robot active vision: A survey of systems, algorithms, and applications. Computational Visual Media, 7(1), 3-29. https://doi.org/10.1007/s41095-020-0179-3

Mendoza, C., Cannelle, B., Barthe, L., & Mellado, N. (2020). Supervised learning of the next-best-view for 3D object reconstruction. Pattern Recognition Letters, 133, 188-195. https://doi.org/10.1016/j.patrec.2020.02.024

Nießner, M., Zollhöfer, M., Izadi, S., & Stamminger, M. (2013). Real-time 3D reconstruction at scale using voxel hashing. ACM Transactions on Graphics (TOG), 32(4), Article 169, 1–11. https://doi.org/10.1145/2508363.2508374

Pirinen, A., & Sminchisescu, C. (2016). Deep Reinforcement Learning for Visual Object Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition (CVPR) Workshops (pp. 341-349). https://doi.org/10.1109/CVPRW.2016.49

Schönberger, J. L., & Frahm, J.-M. (2016). Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4104-4113). https://doi.org/10.1109/CVPR.2016.445

Schönberger, J. L., Zheng, E., Pollefeys, M., & Frahm, J.-M. (2016). Pixelwise View Selection for Unstructured Multi-View Stereo. In European Conference on Computer Vision (ECCV) (pp. 501-518). Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6_31

Schönberger, Johannes Lutz, and Jan-Michael Frahm. Structure-from-Motion Revisited. Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Schönberger, Johannes Lutz, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. European Conference on Computer Vision (ECCV), 2016.